Imperial College London

ICIC Imperial Centre for Inference & Cosmology

Photometric Redshifts for Future Cosmological Galaxy Surveys

Daniel Michael Jones

Astrophysics Group
Department of Physics
Imperial College London

Thesis submitted for the degree of Doctor of Philosophy
Imperial College London
2019

Abstract

Photometric galaxy surveys are a very useful probe that can place constraints on cosmological parameters and test the cosmological model. Future galaxy surveys such as the Large Synoptic Survey Telescope (LSST) will increase the precision of their constraints by observing to fainter magnitudes than previous surveys. While the increased depth of these future surveys will produce higher precision constraints, it will also increase the fraction of sources that overlap with other sources along the line of sight, known as blending. Current methods for dealing with these blended sources involve deblending, separating images of sources into their individual constituents. While this enables their analysis using existing methods designed for unblended sources, this separation makes quantifying and propagating all uncertainties difficult. This thesis presents a different approach, applied to the problem of photometric redshifts. By constructing photometric redshift methods that can infer the redshifts of sources directly from blended data, the associated uncertainties can be easily quantified, a vital step for ensuring the final cosmological constraints of galaxy surveys represent an accurate reflection of our state of knowledge. We first generalise existing Bayesian template-based photometric redshift methods to the case of blended sources. By performing parameter inference on the resulting model, we obtain joint posterior distributions of the redshifts of all constituents within a blended source, completely describing all correlations between these quantities. We then cast the problem of identifying the number of constituents within a blended source as a model comparison problem. Next, we develop a machine learning-based photometric redshift method that can infer the redshifts of sources after being trained on a training set of unblended sources. By using a Gaussian mixture model to do this, the posterior distributions and Bayesian evidences necessary for model comparison can be computed efficiently, enabling the method to be applied to large datasets. Finally, we develop two Bayesian hierarchical models that can infer posterior distributions over redshift distributions of a population of possibly blended sources. We do this by constructing three Gaussian mixture models that share means and covariances but differ in their weights. We test these models using both exact and approximate inference methods. Finally, we conclude by suggesting several possible extensions to this work.

Acknowledgements

It is nigh-on impossible for me to adequately summarise all of the support I have received during my time completing this PhD. Nevertheless, I hope that these words can convey my immense gratitude to the people without whom this thesis would have inevitably never made it into being.

Firstly, I'd like to thank my supervisor Alan. His constant mentorship, guidance and good-humour throughout my time at Imperial have brought me through many ups-and-downs to this point. The way he has encouraged my exploration of ideas with just the right amount of pushing back has irrevocably changed the way I think, and I'm all the better for it. I could not have asked for a better supervisor.

I'd also like to thank everyone in the Imperial astro group for having cultivated such a welcoming place to work. I'd particularly like to thank the other members of my cohort, Tai-an and Sebastian, for having shared this experience with me. I'm very grateful to Andrew and Daniel for their academic input at many points throughout my PhD and, ultimately, for greatly improving the work in this thesis. I'd also like to thank the cosmology group – George, Harry, Wahid, Selim, Lena, Florent, Elena – for broadening my cosmological horizon through our weekly meetings. Finally, a special mention to everyone involved with the planetarium shows – Josh, Tom, Charlotte, Arianna – for the great fun I had.

Next I'd like to thank my family – especially Mum, Dad, Baz, Allie, Chloe, Nan and Gramps – for their endless love and support, without which I could never have completed this PhD. A special mention goes to Cecilia, for her constant company throughout the writing of this thesis. I'd also like to thank my friends – in particular Mike, for being a constant source of fun, encouragement and a wise word.

Finally, I'd like to highlight the immeasurable contribution of my wife Kathryn. Her ceaseless love, care and compassion have given me incalculable strength to get through this process, and her untold patience has improved this thesis in uncountable ways, particularly through her unflinchingly thorough highlighting of my run-on sentences. The four years of this PhD have been both the most wonderful and demanding time of my life so far, and there is no other person I would have rather spent them with than you. *Diolch.*

Bûm gall unwaith – hynny oedd, llefain pan ym ganed.

I was wise once – when I was born, I cried.

The contents of this thesis are the author's own work, except where explicitly indicated.

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution 4.0 International Licence (CC BY). Under this licence, you may copy and redistribute the material in any medium or format for both commercial and non-commercial purposes. You may also create and distribute modified versions of the work. This on the condition that you credit the author. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

Daniel Michael Jones
October 2019

Contents

List of Tables

List of Figures

Part I

Introduction

Chapter 1

Introduction

Cosmology is a science with an incredibly vast history, with most great civilisations of the past having expended some effort in observing the night sky and pondering their existence. However, our modern understanding of cosmology can arguably trace its roots to the observations of Slipher (1917), Lemaître (1927) and Hubble (1929) who discovered that the Universe was expanding.

Cosmology can now be modelled to a very high accuracy through the Λ-cold dark matter (ΛCDM) model, based on Einstein's theory of general relativity (Einstein, 1915) and the Friedmann-Lemaître-Robertson-Walker (FLRW) metric (Friedmann, 1922; Lemaître, 1931; Robertson, 1935; Walker, 1937). This model is now supported by an impressive variety of different cosmological probes, including type-Ia supernovae (e.g., Riess et al., 1998; Perlmutter et al., 1999), the cosmic microwave background (e.g., Planck Collaboration et al., 2018a) and baryon acoustic oscillations (e.g., Eisenstein et al., 2005).

In the past century, cosmology has progressed from a science with very little data to one inundated with it, ushering in the *era of precision cosmology* (e.g., Cortês, 2010; Gerbino, 2014; Akrami et al., 2018). With the launch of future experiments such as the Large Synoptic Survey Telescope (LSST, Ivezić et al., 2019) and the Square Kilometre Array (SKA, Dewdney et al., 2009), this progress in data-volume shows no signs of stopping.

This forthcoming wealth of data will bring opportunity for increases in precision and more careful tests of the cosmological model. However, it will also be the source of a variety of difficulties. Such *big data* may mean that scaling existing analysis methods to future surveys is computationally infeasible, necessitating an increased use of analysis techniques such as machine learning. At the same time, the increased precision of these future surveys only strengthens the need for a complete and accurate understanding

of our uncertainties, as the increased significance of any neglected systematic effects could erroneously point to new physics.

Throughout this thesis, we are concerned with cosmological galaxy surveys, experiments that image galaxies over large volumes of the Universe. To use these surveys to constrain cosmology, it is necessary to know the redshifts of the sources observed. While spectroscopic observations would provide the highest precision measurements of these redshifts, these sources are both too faint and too numerous for spectroscopy to be practically viable. Instead, photometric redshifts must be used, statistical methods that estimate these redshifts from a small number of broadband flux measurements. It is these methods on which the research work of this thesis focusses.

Future cosmological galaxy surveys will increase the precision of their cosmological constraints over current surveys by observing to fainter magnitudes, increasing the number density of sources they observe. One of the consequences of this is that the number of sources that overlap with other sources along the line of sight will greatly increase. This effect is known as blending. In order to use blended sources for cosmology, deblending methods have been developed that separate these sources into their individual constituents. While this allows these sources to be analysed using existing methods designed for unblended sources, it is difficult to correctly account for all uncertainties during this separation.

The research work of this thesis presents a different approach to this problem. Rather than attempting to deblend sources, we develop photometric redshift methods that can identify the redshifts of blended sources using the blended data itself. In this way, all uncertainties associated with the problem, including correlations between each constituent, can be accounted for a propagated further through the cosmological analysis.

The rest of this thesis is structured as follows. The remainder of part I presents the introductory context necessary for the research work of this thesis. Chapter 2 introduces the ΛCDM cosmological model in more detail, and describes various observational probes that can be used to test and constrain it. Chapter 3 discusses various statistical methods that are used throughout this thesis. Chapter 4 details photometric redshift methods for both individual sources and for populations of sources. These methods can be broadly characterised into two types, template-based and empirical methods. This chapter discusses the distinctions between these two types of methods. Finally, chapter 5 discusses the problem of blending in more detail and describes some of the deblending methods that have been developed to address it.

Part II of this thesis presents our original research. Chapter 6 generalises existing

template-based photometric redshift methods to the case of blended sources. We cast the estimation of redshifts as a parameter inference problem, and the identification of the number of constituents in a source as a model comparison problem. We then test our method on both simulated and real flux data.

Chapter 7 tackles the same blended photometric redshift problem using machine leaning techniques, the other main type of photometric redshift method. We use a Gaussian mixture model to fit the joint flux-redshift distribution, and use this model to derive posteriors and Bayesian evidences for one- and two-constituent sources. Our choice of Gaussian mixture model here renders these applications much more computationally efficient than the method in chapter 6, and thus applicable to future cosmological galaxy surveys.

Chapter 8 extends the mixture model approach of chapter 7 to infer posterior distributions over redshift distributions for populations of possibly blended sources. We do this by constructing a Bayesian hierarchical model to infer three independent sets of mixture weights parametrising these distributions.

Finally, part III present our conclusions. Chapter 9 summarises this thesis, and discusses several possible extensions that could be pursued in the future.

Chapter 2

Cosmological Theory and Observations

During the past century, our understanding of cosmology has progressed significantly, from the belief that the Milky Way constituted the entire universe to the present era of precision cosmology. Based on the theory of general relativity devised by Einstein (1915), cosmology can now be described to high precision by a single theory consisting of only 6 parameters, known as Λ-cold dark matter (ΛCDM).

According to this ΛCDM model and our current constraints on its parameters (e.g., Planck Collaboration et al., 2018a), only $\approx 30\%$ of the current energy density of the universe consists of matter. The majority of this matter, corresponding to $\approx 25\%$ of the total energy density, is cold *dark matter* that does not interact electromagnetically. Only the remaining $\approx 5\%$ of the energy density is baryonic. Instead of matter, the vast majority of the energy density of the universe, corresponding to $\approx 70\%$, consists of *dark energy*, a poorly understood energy driving the accelerated expansion of the universe.

The theoretical basis for ΛCDM is discussed in more detail in section 2.1. Despite many observational successes of ΛCDM, some questions still remain. The nature of dark matter and dark energy are poorly understood, and it is possible that these could be described instead by a modification to gravity. In addition, some tensions still exist between different cosmological probes measuring the same quantities. Section 2.2 describes some common cosmological probes that can be used to investigate these. Finally, section 2.3 discusses cosmological galaxy surveys in more detail, another type of cosmological probe that can be used to constrain and test cosmological models. It is these photometric galaxy surveys that the photometric redshift methods developed in the research part of this thesis are applicable to.

2.1 The ΛCDM Model of Cosmology

This section briefly introduces the ΛCDM model of cosmology, discussing its basis in the theory of general relativity and some of its theoretical predictions.

2.1.1 Hubble's law and the expansion of the universe

It is now well established by cosmological observations that the universe is expanding. This was first seen in the recessional velocity of galaxies as was observed by Slipher (1917), Lemaître (1927) and Hubble (1929), where galaxies at greater distances from us have greater recessional velocities. There is a linear relation between these quantities known as Hubble's law[1], given by

$$v = H_0 d\,, \tag{2.1}$$

where v is the recessional velocity of the galaxy, d is its proper distance, and H_0 is the constant of proportionality, known as the Hubble constant.

The Hubble constant H_0 is the *current* rate of expansion, though in general this rate is a time-dependent quantity. This rate is given by the Hubble parameter, defined as

$$H(t) = \frac{\dot{a}(t)}{a(t)}\,, \tag{2.2}$$

where $a(t)$ is the scale factor, and $\dot{a}(t)$ denotes its time derivative. The scale factor is a dimensionless quantity that describes the relative size of the universe, and is normalised such that

$$a_0 \equiv a(t_0) = 1\,, \tag{2.3}$$

where t_0 is the current time. The scale factor is used to convert between proper distances d and comoving distances χ, i.e.,

$$d = a(t)\chi\,. \tag{2.4}$$

By being defined in this way, comoving distances are invariant under expansions of the universe.

[1]This can also be referred to as the Hubble-Lemaître law to acknowledge Lemaître's contribution to this discovery, oft-forgotten due to the removal of key sections in the original english translation of his work (Livio, 2011).

2.1.2 The cosmological principle

The underlying assumption of modern cosmology is the *cosmological principle*, which states that the universe is both homogeneous and isotropic on sufficiently large scales. Homogeneity means the statistics of the universe are invariant to translation, i.e., the universe looks the same in all places. Isotropy means that the statistics of the universe are invariant to rotation, i.e., the universe looks the same in all directions. Broadly speaking, the cosmological principle can be taken to mean that we are not privileged observers of the universe, but rather that our view of the universe is typical and representative.

2.1.3 General relativity

On cosmological scales, gravity is the dominant force that drives the dynamics and evolution of the universe. In the standard ΛCDM cosmological model, this is assumed to be described by general relativity, a geometric theory of gravity where the force is described by curvature of spacetime. The motions of objects through this spacetime are influenced by this curvature, and the presence of these objects affects the curvature of the spacetime. This relationship is summarised by Einstein's field equations (Einstein, 1917), given by

$$R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu} + \Lambda g_{\mu\nu} = \frac{8\pi G}{c^4}T_{\mu\nu}\,, \tag{2.5}$$

where c is the speed of light, and G is the gravitational constant. We discuss the other terms in this equation below.

The term on the right-hand side of Einstein's field equations is the stress-energy tensor $T_{\mu\nu}$. This term describes the energy content of the universe in 16 components when in $3+1$-dimensional spacetime, though since this tensor is symmetric $T_{\mu\nu} = T_{\nu\mu}$, only 10 of these components are free.

The spacetime curvature is encapsulated within $g_{\mu\nu}$, the metric tensor. The metric of a flat spacetime is known as the Minkowski metric $g_{\mu\nu} = \eta_{\mu\nu}$, given by

$$\eta_{\mu\nu} = \begin{bmatrix} -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{2.6}$$

The metric defines the line element

$$\mathrm{d}s^2 = g_{\mu\nu}\,\mathrm{d}X^\mu\,\mathrm{d}X^\nu\,, \tag{2.7}$$

where we make use of the Einstein summation convention. This is the invariant distance between two points in spacetime, i.e., its value is agreed upon by all observers.

Also in equation 2.5 are $R_{\mu\nu}$ and R, the Ricci tensor and Ricci scalar respectively. Their combination in the field equations is often summarised in the Einstein tensor $G_{\mu\nu}$, given by

$$G_{\mu\nu} = R_{\mu\nu} - \frac{1}{2}Rg_{\mu\nu}\,. \tag{2.8}$$

The left-hand side of the Einstein field equations defines the curvature of the spacetime. The Ricci tensor and Ricci scalar defining the Einstein tensor can both be calculated as a function of the metric $g_{\mu\nu}$. Thus, the geometry of the spacetime can be specified completely through the metric.

The final term in equation 2.5 is the cosmological constant Λ, an optional term within Einstein's field equations. This term was introduced by Einstein to ensure the prediction of a static universe, a decision he described as his "biggest blunder" once observational evidence indicated that the universe was expanding (Gamow, 1956; O'Raifeartaigh and Mitton, 2018). Ironically, this term is now seen as a possible description of the *accelerated* expansion of the universe (e.g., Peebles and Ratra, 2003) that was discovered by observations of type Ia supernovae (Riess et al., 1998; Perlmutter et al., 1999). This is discussed in more detail in section 2.1.8.

The terms on the left-hand side of the Einstein field equations can be seen to describe *intrinsic* properties of the spacetime, such as its curvature. This is in contrast to terms on the right-hand side, which describe the *contents* of the spacetime. We note that the convention shown here of writing the cosmological constant on the left-hand side therefore implies that it is a property of the spacetime itself, rather than a manifestation of something contained within it, such as an additional field.

2.1.4 The Friedmann-Lemaître-Robertson-Walker metric

Solutions to Einstein's field equations are specified as a metric. The Friedmann-Lemaître-Robertson-Walker (FLRW)[2] metric (Friedmann, 1922; Lemaître, 1931; Robertson, 1935; Walker, 1937) describes an expanding universe that is homogeneous and isotropic, as described above, and therefore forms the basis of the ΛCDM cosmological model. This metric, defined in comoving polar coordinates (r, θ, ϕ), is given by

$$\mathrm{d}s^2 = -c^2\,\mathrm{d}t^2 + a^2(t)\left(\frac{\mathrm{d}r^2}{1-Kr^2} + r^2\,\mathrm{d}\theta^2 + r^2\sin^2\theta\,\mathrm{d}\phi^2\right)\,, \tag{2.9}$$

[2]We use the full initialism throughout this thesis, though FRW is also commonly used.

where K is the curvature describing the geometry of the universe. When $K = 0$, the universe is flat. In this case, the universe would have an *Euclidean* geometry, meaning that parallel lines remain parallel when extended to infinity, and the internal angles of a triangle sum to 180°. When the curvature $K \neq 0$, the universe has a non-Euclidean geometry. A universe with positive curvature $K > 0$ is said to be closed. This universe would have a spherical geometry, meaning that parallel lines intersect when extended and the internal angles of a triangle sum to $> 180°$. A universe with negative curvature $K < 0$ is said to be open. This universe would have a hyperbolic geometry, meaning that parallel lines diverge when extended and the internal angles of a triangle sum to $< 180°$. Observations of the cosmic microwave background, detailed in section 2.2.1, indicate that our universe is consistent with zero curvature and would thus be flat (Planck Collaboration et al., 2018a).

2.1.5 Cosmological redshift

Redshift is an effect where the observed wavelength of light differs from that of when it was emitted. This is quantified in the dimensionless value z, defined as

$$z = \frac{\lambda_o - \lambda_e}{\lambda_e} \tag{2.10}$$

where λ_e is the wavelength of the light when it is emitted λ_o is the wavelength of the light when it is observed. In non-cosmological settings, this effect can arise as a Doppler shift due to the source of the light having a velocity relative to the observer. However, in cosmological settings, redshift arises not due to a recessional velocity, but rather as the result of photons propagating through an expanding universe. It is predominantly this effect[3] which we measure throughout this thesis. We derive this effect below, following the exposition of Theuns (2016).

We first note that light travels along null geodesics, i.e., paths where $\mathrm{d}s^2 = 0$. Consider a *radial* light ray, propagating along a null geodesic in a coordinate system orientated such that $\mathrm{d}\theta = \mathrm{d}\phi = 0$ with the observer at the origin. In this case, equation 2.9 can be rearranged to give

$$\frac{c}{a(t)}\,\mathrm{d}t = \frac{\mathrm{d}r}{\sqrt{1 - Kr^2}}\,. \tag{2.11}$$

[3]The total observed redshift of a galaxy also includes a Doppler shift contribution due to its local velocity relative to the Hubble flow. However, this is only non-negligible at very small redshifts.

Integrating both sides, this defines the comoving distance χ as

$$\chi \equiv \int_{t_e}^{t_o} \frac{c}{a(t)}\,\mathrm{d}t = \int_0^{r_e} \frac{\mathrm{d}r}{\sqrt{1-Kr^2}}\,. \tag{2.12}$$

where r_e is the radial comoving coordinate of a source that emits a photon at time t_e which is observed at time t_o. This comoving distance is invariant as the universe expands, provided that the source and observer are both stationary in comoving coordinates. As a result, χ will be equal for a second photon emitted a short time δ_e later. This photon will then be observed a short time δ_o later, where $\delta_e \neq \delta_o$ due to the expansion of the universe. Thus,

$$\int_{t_e}^{t_o} \frac{c}{a(t)}\,\mathrm{d}t = \int_{t_e+\delta_e}^{t_o+\delta_o} \frac{c}{a(t)}\,\mathrm{d}t\,. \tag{2.13}$$

Note that the integral on the right-hand side can be written as a sum of integrals over different intervals, i.e.,

$$\int_{t_e+\delta_e}^{t_o+\delta_o} \frac{c}{a(t)}\,\mathrm{d}t = \int_{t_e}^{t_o} \frac{c}{a(t)}\,\mathrm{d}t + \int_{t_o}^{t_o+\delta_o} \frac{c}{a(t)}\,\mathrm{d}t - \int_{t_e}^{t_e+\delta_e} \frac{c}{a(t)}\,\mathrm{d}t\,. \tag{2.14}$$

Thus,

$$\int_{t_o}^{t_o+\delta_o} \frac{c}{a(t)}\,\mathrm{d}t = \int_{t_e}^{t_e+\delta_e} \frac{c}{a(t)}\,\mathrm{d}t\,. \tag{2.15}$$

Finally, assuming that the time intervals are small in comparison with timescales for changes in $a(t)$, these integrals can be approximated to give

$$\frac{c\delta_o}{a(t_0)} = \frac{c\delta_e}{a(t_e)}\,. \tag{2.16}$$

This can then be written in terms of the wavelengths λ_e and λ_o as

$$\frac{a(t_0)}{a(t_e)} = \frac{\lambda_o}{\lambda_e}\,. \tag{2.17}$$

Thus, inserting the definition of redshift from equation 2.10, the cosmological redshift can be related to the scale factor by

$$\frac{1}{a} = 1 + z \tag{2.18}$$

where $a \equiv a(t_e)$, and we use the fact that the scale factor is defined so that its present-day value is unity, i.e., $a(t_0) = 1$.

2.1.6 The Friedmann Equations

By inserting the FLRW metric into Einstein's field equations in equation 2.5, it is possible to derive two equations describing the dynamics of the scale factor a, known as the Friedmann equations. These are given by

$$H^2 = \left(\frac{\dot{a}}{a}\right)^2 = \frac{8\pi G}{3}\rho - \frac{Kc^2}{a^2} + \frac{\Lambda}{3} \tag{2.19}$$

and

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left(\rho + \frac{3p}{c^2}\right) + \frac{\Lambda}{3}\,, \tag{2.20}$$

where c and G are the speed of light and the gravitational constant, K is the curvature as defined above, ρ is the total energy density and p is the pressure. The two Friedmann equations above can be used to derive the conservation equation

$$\dot{\rho} + 3H\left(\rho + \frac{p}{c^2}\right) = 0 \tag{2.21}$$

The content of the universe is modelled as a perfect fluid, where the pressure and the energy density are assumed to be linearly related, i.e.,

$$p = w\rho c^2\,. \tag{2.22}$$

This is known as the *equation of state*, with w being the equation of state parameter. This parameter $w = 1/3$ for radiation and $w = 0$ for non-relativistic matter.

Energy densities

By inserting the equation of state into the conservation equation defined in equation 2.21, it becomes

$$\frac{\dot{\rho}}{\rho} = -3\,(1+w)\,\frac{\dot{a}}{a}\,, \tag{2.23}$$

where we have inserted the definition of the Hubble parameter from equation 2.2. The solution to this equation is given by

$$\rho \propto a^{-3(1+w)} \tag{2.24}$$

Inserting the values of the equation of state parameter w from above and using the fact the $a_0 = 1$, we find that the energy density of radiation scales as

$$\rho_r = \rho_{r,0}a^{-4}\,, \tag{2.25}$$

and the energy density of non-relativistic matter scales as

$$\rho_m = \rho_{m,0} a^{-3} \,, \tag{2.26}$$

where $\rho_{r,0}$ and $\rho_{m,0}$ refer to the present-day values of the radiation and matter energy density respectively. It is common to work with dimensionless density parameters that are defined relative to the critical density ρ_c. This critical density is defined as the energy density that, in a universe without a cosmological constant, i.e., $\Lambda = 0$, there is zero curvature $K = 0$ and the universe is flat. Inserting these into the first Friedmann equation defined in equation 2.19, this critical density is given by

$$\rho_c = \frac{3H^2}{8\pi G} \,. \tag{2.27}$$

The density parameters are then defined as ratios with respect to this critical value. The radiation density parameter is given by

$$\Omega_r = \frac{\rho_r}{\rho_c} \,, \tag{2.28}$$

and the matter density parameter is given by

$$\Omega_m = \frac{\rho_m}{\rho_c} \,. \tag{2.29}$$

We can also define an analogous density parameter corresponding to the curvature, given by

$$\Omega_K = -\frac{Kc^2}{H^2 a^2} \,. \tag{2.30}$$

Evaluating this with the present-day values of the Hubble parameter and the scale factor defines the present-day density parameter $\Omega_{K,0}$. Finally, we can also define the cosmological constant density parameter, given by

$$\Omega_\Lambda = \frac{\Lambda}{3H^2} \,. \tag{2.31}$$

As before, evaluating this using the present-day values of the Hubble constant gives $\Omega_{\Lambda,0}$. The cosmological constant can also be thought of as a fluid with equation of state parameter $w = -1$, so that

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} \,, \tag{2.32}$$

where ρ_Λ is constant. While ρ_Λ is constant, ρ_c is redshift dependent, since it is defined in terms of the Hubble parameter. As a result, the density parameter of the cosmological constant Ω_Λ evolves with redshift. We detail how these parameters evolve with redshift

below.

Evolution with redshift

Using the density parameters above, the Hubble parameter can be written as a function of redshift, given by

$$H = H_0\sqrt{\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{K,0}(1+z)^2 + \Omega_{\Lambda,0}}\,, \tag{2.33}$$

where $\Omega_{r,0}$, $\Omega_{m,0}$, $\Omega_{\Lambda,0}$ and $\Omega_{K,0}$ are the present-day radiation, matter, cosmological constant and curvature density parameters. The terms under the square root are commonly written as the function

$$E(z) = \sqrt{\Omega_{r,0}(1+z)^4 + \Omega_{m,0}(1+z)^3 + \Omega_{k,0}(1+z)^2 + \Omega_{\Lambda,0}}\,, \tag{2.34}$$

so that

$$H = H_0 E(z)\,. \tag{2.35}$$

We can also write the density parameters in terms of the redshift using this function. These are given by

$$\Omega_r = \frac{\Omega_{r,0}(1+z)^4}{E^2(z)} \tag{2.36}$$

for the radiation,

$$\Omega_m = \frac{\Omega_{m,0}(1+z)^3}{E^2(z)} \tag{2.37}$$

for the matter, and

$$\Omega_\Lambda = \frac{\Omega_{\Lambda,0}}{E^2(z)}\,. \tag{2.38}$$

for the cosmological constant. It is also possible to find the redshift of equality when the densities of matter and radiation were equal by equating the density parameters above to give

$$\Omega_{r,0}(1+z_{eq}^{r,m})^4 = \Omega_{m,0}(1+z_{eq}^{r,m})^3\,, \tag{2.39}$$

where we label the redshift of equality of radiation and matter by $z_{eq}^{r,m}$ Thus,

$$z_{eq}^{r,m} = \frac{\Omega_{m,0}}{\Omega_{r,0}} - 1\,. \tag{2.40}$$

Similarly for matter and the cosmological constant,

$$z_{eq}^{m,\Lambda} = \sqrt[3]{\frac{\Omega_{\Lambda,0}}{\Omega_{m,0}}} - 1\,. \tag{2.41}$$

By inferring the present-day values of the density parameters $\Omega_{r,0}$, $\Omega_{m,0}$ and $\Omega_{\Lambda,0}$, the evolution of the energy content of the universe can be predicted. As described in section 2.2.1, these values can be inferred using observations of the cosmic microwave background, as was done by the Planck mission (Planck Collaboration et al., 2018b). The most recent data release presented in Planck Collaboration et al. (2018a) finds[4] good agreement with a flat universe $\Omega_{K,0} = 0$, where $\Omega_{\Lambda,0} = 0.6847 \pm 0.0073$ and $\Omega_{m,0} = 0.3153 \pm 0.0073$. The redshift of radiation and matter equality is found to be $z_{eq}^{r,m} = 3402 \pm 26$. Thus, using equation 2.40, we can calculate that $\Omega_{r,0} = 9.265 \times 10^{-5}$.

These results indicate that only a negligible proportion of the total energy budget of the universe is given by radiation, while matter makes up a larger but still minority contribution. Instead, the current energy budget of the universe is dominated by the cosmological constant. However, this was not always the case. Using the expressions defined above and the present-day cosmological parameters given above, we can predict the evolution of these contributions. This is shown in Figure 2.1.

This figure shows that there have been three energy content epochs throughout the history of the universe. Firstly, the early universe was radiation dominated. Next, matter became the dominant contribution, passing radiation at the redshift of matter-radiation equality $z_{eq}^{r,m}$. As indicated above, the results from Planck indicate that $z_{eq}^{r,m} = 3402 \pm 26$. Finally, the present-day universe is dominated by the cosmological constant. From equation 2.41 and the Planck cosmological parameters, we can estimate the redshift this began to be $z_{eq}^{m,\Lambda} = 0.295$. The consequences of the universe currently being dominated by the cosmological constant are discussed in section 2.1.8.

2.1.7 Cosmological distance measures

We now come to define several distance measures used in cosmology, following the definitions presented in Hogg (1999). The first is comoving distance χ. Like the comoving coordinates in the FLRW metric defined in equation 2.9, this distance measure is invariant under the expansion of the universe. As in equation 2.12, this can be written as an integral over time as

$$\chi = \int_{t_e}^{t_o} \frac{c}{a(t)} \, \mathrm{d}t \,, \tag{2.42}$$

where χ is the comoving distance travelled by a light ray emitted at time t_e and observed at time t_o. However, this comoving distance can also be written in terms of redshift as

$$\chi = \frac{c}{H_0} \int_0^z \frac{\mathrm{d}z'}{E(z')} \,, \tag{2.43}$$

[4]We use the TT,TE,EE+lowE+lensing values from table 2.

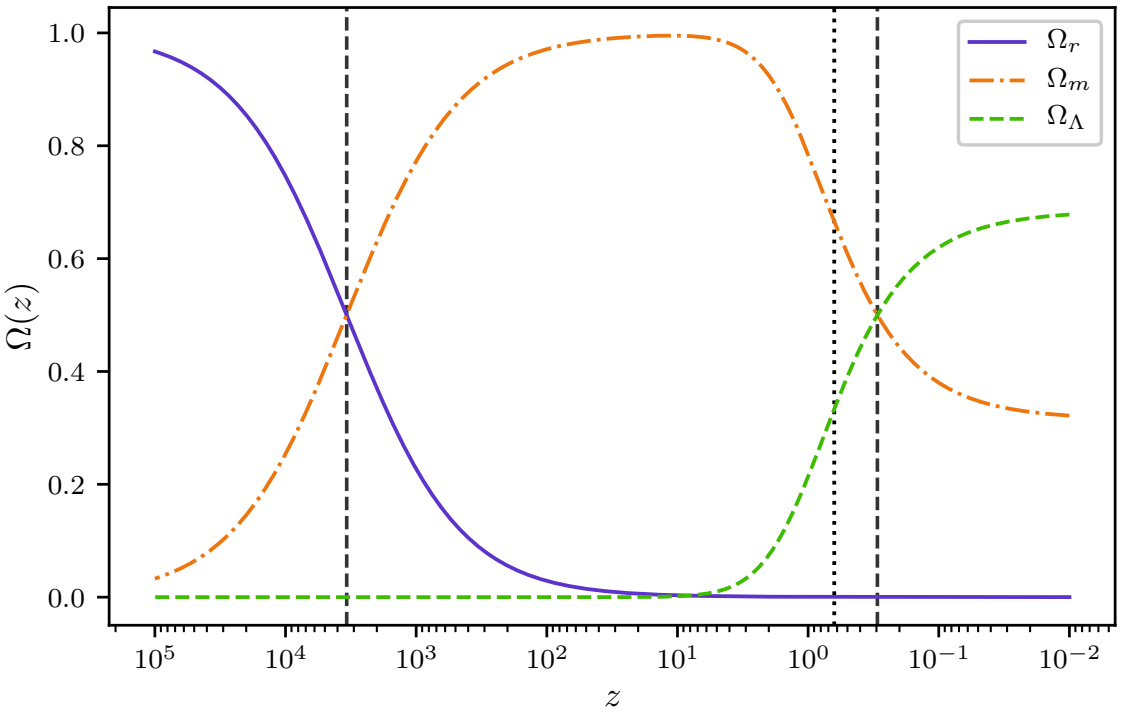

Figure 2.1: Plot showing the evolution of the density parameters with redshift, calculated assuming a flat Planck (Planck Collaboration et al., 2018a) cosmology. The solid purple line shows the radiation density parameter Ω_r, the dot-dashed orange line shows the matter density parameter Ω_m, and the dashed green line shows the cosmological constant density parameter Ω_Λ. The vertical dashed lines indicate the redshifts of matter-radiation equality $z_{eq}^{r,m} = 3402$ and matter-Λ equality $z_{eq}^{m,\Lambda} = 0.295$. The vertical dotted line at $z \approx 0.631$ shows the redshift where acceleration began, derived in section 2.1.8. The redshifts on the x-axis are plotted in reverse so that the present-day is towards the right-hand side of the plot.

where χ is the comoving distance out to redshift z from an observer at redshift zero, i.e., on Earth. The function $E(z)$ is defined as in equation 2.34.

A quantity which is closely related to the comoving distance is proper distance, the distance that would be measured with a ruler at a specific point in time. This is related to the comoving distance through the scale factor as in equation 2.4.

The comoving distance above is a *line of sight* distance; that is, the distance measured along the path of a propagating light ray from a source at redshift z to the observer. We can also define the comoving distance between two sources at the same redshift z but separated by an angle θ on the sky. This is given by θD_m, where D_m is the *transverse* comoving distance, given by

$$D_m = \begin{cases} \frac{c}{H_0\sqrt{|\Omega_K|}} \sin\left(\frac{\chi H_0\sqrt{|\Omega_K|}}{c}\right) & \text{for } \Omega_K < 0 \\ \chi & \text{for } \Omega_K = 0 \\ \frac{c}{H_0\sqrt{\Omega_K}} \sinh\left(\frac{\chi H_0\sqrt{\Omega_K}}{c}\right) & \text{for } \Omega_K > 0\,. \end{cases} \tag{2.44}$$

When in a Minkowski spacetime, a source which emits light with a total luminosity L can be observed at a distance d to have a flux F given by

$$F = \frac{L}{4\pi d^2}\,. \tag{2.45}$$

The *luminosity distance* D_L generalises this notion, and so is defined as

$$D_L = \sqrt{\frac{L}{4\pi F}}\,. \tag{2.46}$$

This distance can be calculated from the transverse comoving distance D_m defined above by

$$D_L = (1+z)D_m\,. \tag{2.47}$$

The luminosity distance can also be calculated from equation 2.46 by observing sources with a known luminosity. Such sources are known as *standard candles*. An example of a standard candle is a type Ia supernovae[5]; these are discussed in more detail in section 2.2.2.

Finally, we consider *angular diameter distance* D_A. This distance is defined from

[5]In general, some sources may not have a standard luminosity, but can instead have their luminosity calibrated using another observable feature, such as the change in flux with time. These sources are then referred to as *standardisable* candles. Type Ia supernovae are standardisable in this way.

the physical size of a source r and the angle it subtends on the sky θ by

$$D_A = \frac{r}{\theta}\,. \tag{2.48}$$

Analogously to calculating the luminosity distance above using standard candles with a known luminosity, the angular diameter distance can be calculated using this definition from *standard rulers*, objects with a known length. Baryon acoustic oscillations are an examples of such an object; these are discussed in more detail in section 2.2.3. The angular diameter distance is also related to the transverse comoving distance by

$$D_A = \frac{D_m}{(1+z)}\,, \tag{2.49}$$

and so to the luminosity distance by

$$D_A = \frac{D_L}{(1+z)^2}\,. \tag{2.50}$$

Figure 2.2 shows a plot of these distances against redshift for a flat universe with cosmological parameters given by Planck (Planck Collaboration et al., 2018a). In addition to the density parameters given in section 2.1.6, we need the Hubble constant to calculate these distances. This was found by Planck[6] to be $H_0 = 67.36 \pm 0.54$ km s^{-1} Mpc^{-1}. The most striking feature of this figure is that the angular diameter distance does not increase indefinitely. Instead, D_A decreases with redshift at redshifts of $z \gtrsim 1.5$. Thus, due to the definition of the angular diameter distance in equation 2.48, the angular size of a source with a fixed physical size will *increase* as the source is placed more distantly.

2.1.8 Accelerated expansion of the universe

While it has been understood for nearly a century that the universe is expanding, the expectation was that this expansion would be slowing down. On cosmological scales, gravity acts to pull massive objects, such as galaxies, closer together. This gravitational pull would therefore be expected to act against the expansion of the universe, slowing it down and perhaps eventually reversing the direction so that the universe began to contract. However, more recent observational evidence has not borne out this expectation. As first indicated by observations of type-Ia supernovae by Riess et al. (1998) and Perlmutter et al. (1999), the expansion of the universe is not decelerating,

[6]The value of the Hubble constant found by Planck is in tension with that determined by local measurements. This is discussed in more detail in section 2.3.4.

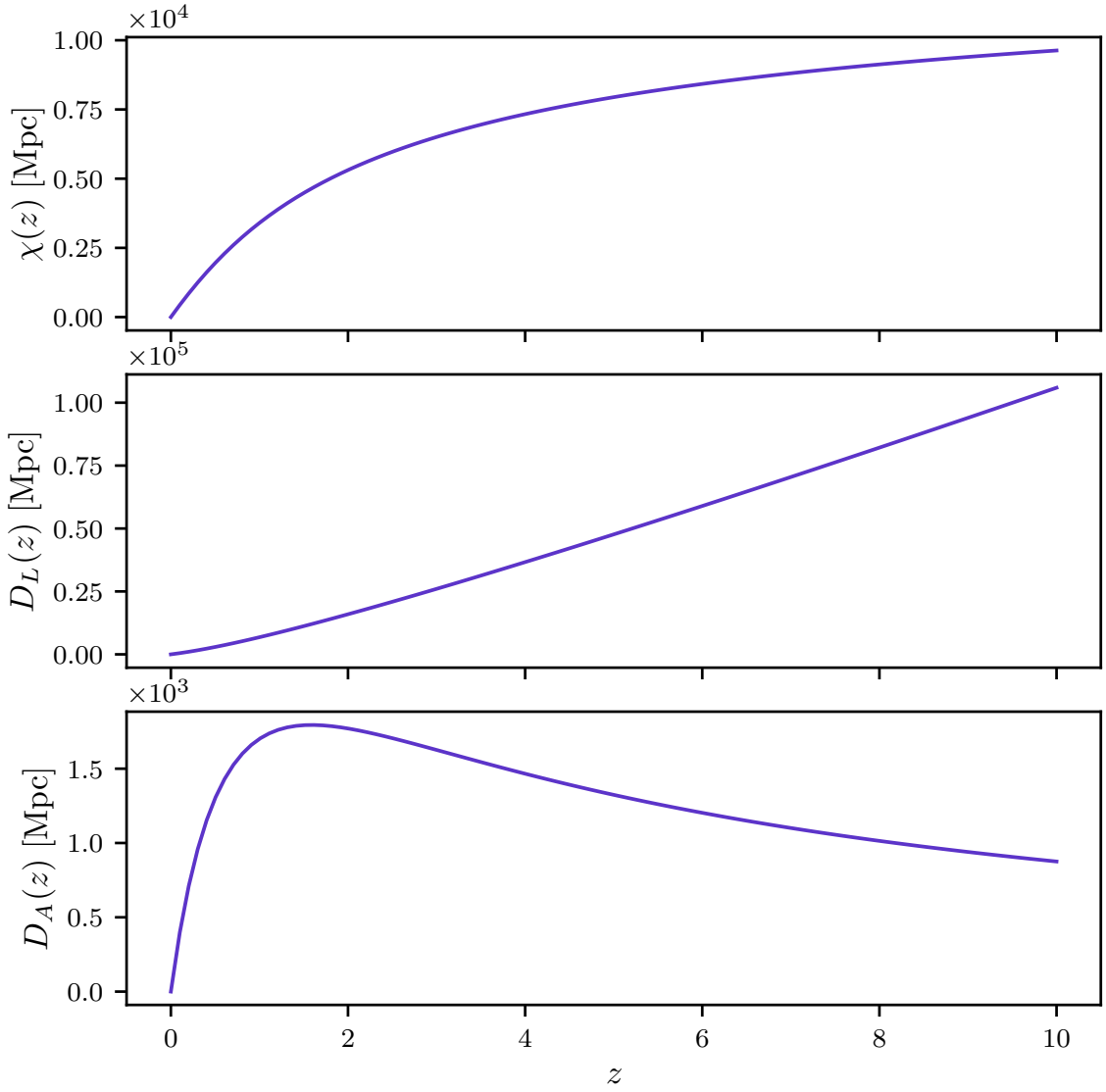

Figure 2.2: Plot showing various cosmological distance measures vs redshift z, calculated assuming a flat Planck (Planck Collaboration et al., 2018a) cosmology. The top panel shows the comoving distance $\chi(z)$, which equals the transverse comoving distance D_m since we assume a flat cosmology. The middle panel shows the luminosity distance D_L. The bottom panel shows the angular diameter distance D_A. Note that unlike the other distance measures, D_A decreases at redshifts $z \gtrsim 1$.

but rather is *accelerating*. These observations were later the subject of the 2011 Nobel prize as a result[7]. The observations of these supernovae are discussed in more detail in section 2.2.2

A source of energy is necessary to drive this accelerated expansion; this is known as *dark energy*. The cosmological constant Λ presents a possible description of this dark energy that is consistent with current observations. Given the value of $\Omega_{\Lambda,0} = 0.6847 \pm 0.0073$ from Planck (Planck Collaboration et al., 2018a) above, ΛCDM then predicts this accelerated expansion. To see this, we can model the cosmological constant as a fluid in the same way as we did with radiation and matter. We detail this below, following the explanation by Liddle (2003).

By modelling the cosmological constant as a fluid, we are therefore able to write a corresponding energy density ρ_Λ. As with radiation and density, the density parameter Ω_Λ can be defined in terms of the energy density ρ_Λ using the critical density ρ_c as

$$\Omega_\Lambda = \frac{\rho_\Lambda}{\rho_c} \,. \tag{2.51}$$

By using the definitions of Ω_Λ from equation 2.31 and ρ_c from equation 2.27, we can see that

$$\rho_\Lambda = \frac{\Lambda}{8\pi G} \,, \tag{2.52}$$

i.e., also a constant. We can now write the conservation equation defined in equation 2.21 for the Λ, giving

$$\dot{\rho}_\Lambda + 3H\left(\rho_\Lambda + \frac{p_\Lambda}{c^2}\right) = 0 \tag{2.53}$$

Since the energy density is constant, $\dot{\rho}_\Lambda = 0$. As a result, the pressure must be related to the density be

$$p_\Lambda = -\rho_\Lambda c^2 \,. \tag{2.54}$$

This is simply the equation of state defined in equation 2.22. Thus, the cosmological constant has an equation of state parameter given by $w = -1$. We can rewrite the second Friedmann equation defined in equation 2.20 in terms of the equation of state parameter as

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\rho\,(1 + 3w) \,, \tag{2.55}$$

where Λ no longer appears due to being absorbed into the energy density term. From this equation, we can then see that $\ddot{a} > 0$ when $w < -1/3$. Since the cosmological constant has an equation of state parameter of $w = -1$, a universe dominated by Λ is

[7]The Nobel Prize 2011 - Press release - `https://www.nobelprize.org/prizes/physics/2011/press-release`

predicted to experience an accelerated expansion.

We can also use equation 2.55 to predict the redshift at which accelerated expansion began. When considering a general epoch of the universe rather than one where a single component dominates, we can sum over all components such that

$$\frac{\ddot{a}}{a} = -\frac{4\pi G}{3}\left[\rho_m\left(1+3w_m\right) + \rho_r\left(1+3w_r\right) + \rho_\Lambda\left(1+3w_\Lambda\right)\right], \tag{2.56}$$

where we use the indices m, r and Λ to refer to matter, radiation and the cosmological constant respectively. Inserting the values of the equation of state parameters and the definition of the critical density from above, we find

$$\frac{\ddot{a}}{a} = -\frac{H^2}{2}\left(\Omega_m + 2\Omega_r - 2\Omega_\Lambda\right). \tag{2.57}$$

Thus, the redshift where the universe began accelerating can be found by solving

$$\Omega_m + 2\Omega_r - 2\Omega_\Lambda = 0\,. \tag{2.58}$$

where the redshift evolution of the density parameters is given in section 2.1.6. Solving numerically for redshift assuming a flat Planck (Planck Collaboration et al., 2018a) cosmology, we find the redshift that acceleration started to be $z \approx 0.631$. We note that this is at an earlier time than matter-Λ equality $z_{eq}^{m,\Lambda} = 0.295$, when the cosmological constant became the dominant contribution.

Alternatives to the cosmological constant

As described in section 2.1.8, the ΛCDM model allows the accelerated expansion of the universe can be modelled and explained through the cosmological constant Λ. General relativity, the model of gravity on which ΛCDM is based, has survived many observational tests.

The measurement of the deflection of light rays through gravitational lensing by the sun during the total solar eclipse of 1919 (Dyson et al., 1920) gave the first observational evidence for a prediction of general relativity. This, alongside the retrodiction of the perihelion precession of Mercury (Einstein, 1916) and other observational predictions such as gravitational redshift (Pound and Rebka, 1959), the Shapiro time delay (Shapiro, 1964) and indirect evidence of gravitational waves from measurements of a binary pulsar system (Hulse and Taylor, 1975; Taylor and Weisberg, 1982; Weisberg et al., 2010) have provided a strong observational basis for general relativity. More recent direct detections of gravitational waves, a prediction of general relativity, have

been made by extremely sensitive laser interferometers (Abbott et al., 2016). This has enabled additional observational tests (e.g., Abbott et al., 2019), all of which have been consistent with general relativistic predictions.

Despite these successes, the cosmological constant has problems from the perspective of fundamental physics. The *cosmological constant problem* is that predictions from quantum field theory of the vacuum energy density, which should correspond with the cosmological constant, are larger than the value inferred from cosmological observations by 120 orders of magnitude[8] (e.g., Carroll, 2002). The substantial difference between prediction and observation has therefore lead to the development of several competing explanations for the accelerated expansion of the universe.

One such explanation within general relativity is that an as-yet-unknown scalar field could be an explanation for dark energy. This is known as *quintessence*. If this scalar field were to have an equation of state parameter $w < -1/3$, this could cause the accelerated expansion as described above. Such theories predict a varying equation of state w, which can also be interpreted as an effective Λ which varies with cosmic time (Solà and Štefančić, 2005).

A simple phenomenological parametrisation for dark energy with a varying equation of state is given by (Chevallier and Polarski, 2001; Linder, 2003)

$$w(a) = w_0 + w_a(1 - a)\,, \tag{2.59}$$

where a is the scale factor, and w_0 and w_a are constants. The values $w_0 = -1$ and $w_a = 0$ corresponds to the cosmological constant, while $w_a \neq 0$ results in a varying equation of state. Constraints on these parameters from Planck (Planck Collaboration et al., 2018a) were found to be compatible with the ΛCDM prediction.

An alternative explanation would be a modification to general relativity, a scenario referred to as *modified gravity*. Many varieties of modified theories of gravity exist. One family of these is $f(R)$ gravity (Buchdahl, 1970), where each theory corresponds to a particular function $f(R)$ of the Ricci scalar R, discussed in section 2.1.3. General relativity forms part of this family through the function $f(R) = R$ (e.g., Sotiriou and Faraoni, 2010). Another large family of modified gravity theories are Horndeski theories (Horndeski, 1974), where modifications are made through an additional scalar field. This family of theories contains both general relativity and quintessence, in addition to many common extensions (e.g., Kobayashi, 2019).

We note that these are simply two large families of modified gravity theories and

[8]This has been described as *"probably the worst theoretical prediction in the history of physics"* (Hobson et al., 2006)

that many more modifications exist. A comprehensive review of these is beyond the scope of this thesis; we instead refer to the reviews of Clifton et al. (2012), Koyama (2016) and Nojiri et al. (2017) for more thorough discussions.

2.2 Observational Probes

The rapid development of cosmology over the previous century has been enabled by access to a wealth of observations. Combining evidence from multiple cosmological probes allows us to break degeneracies associated with each method and ultimately increase the precision on cosmological parameter constraints. These combinations also act as useful cross-checks, with tensions between independent probes potentially pointing to the existence of new physics. In this section, we discuss several methods for making these cosmological observations, and how they are used to constrain the cosmological model and its parameters.

2.2.1 The cosmic microwave background

The early universe consisted of a very hot plasma of protons, electrons and photons. The temperature was sufficiently high that no hydrogen atoms could form, since they would be ionised by the surrounding high-energy photons. As a result, this primordial plasma had many free electrons. The surrounding high energy photons were readily scattered by these free electrons through Thomson scattering, making their mean free path very short. Thus, photons could not freely propagate during this time.

The continued expansion of the universe caused it to cool. Photons that were initially high-energy redshifted and became less energetic. As a result, these photons were no longer able to ionise hydrogen atoms, and so electrons and protons combined to form neutral hydrogen, a process known as *recombination*. Finally, this reduction in the number of free electrons allowed protons to propagate freely, known as *decoupling*. As these photons propagated, they continued to be redshifted, and have microwave wavelengths in the present day. This first light is therefore known as the *cosmic microwave background* (CMB).

The present-day spectrum of the CMB is a blackbody with a temperature of $\approx 2.7\,\mathrm{K}$ (Fixsen et al., 1994). This spectrum has been a blackbody throughout its history, as redshifting a blackbody spectrum results in another blackbody spectrum of a different temperature (Liddle, 2003). This temperature is also very isotropic, with variations limited to 1 part in 10^5 (e.g., Clements, 2017). However, these small

anisotropies are vital for making the CMB a useful cosmological probe. Due to the coupling of photons and matter, small anisotropies in the matter density of the early universe manifested in anisotropies in the temperature of the CMB. Observations of these variations are therefore a probe of the density distribution of the early universe.

The angular power spectrum of these anisotropies can be predicted from cosmological models, allowing observations of the CMB to place constraints on cosmological parameters. We now detail this power spectrum following Hivon et al. (2002). We first describe the temperature of the CMB as a small perturbation over a background average temperature, i.e.,

$$T(\boldsymbol{n}) = \bar{T} + \Delta T(\boldsymbol{n}) \tag{2.60}$$

where $T(\boldsymbol{n})$ is the temperature in direction $\boldsymbol{n}$ on the sky, $\Delta T(\boldsymbol{n})$ is the temperature perturbation in direction $\boldsymbol{n}$, and $\bar{T}$ is the average temperature. The temperature perturbations $\Delta T(\boldsymbol{n})$ are the anisotropies we wish to define the angular power spectrum of. These anisotropies are then decomposed into spherical harmonics $Y_{\ell m}$ by

$$\Delta T(\boldsymbol{n}) = \sum_{\ell>0} \sum_{m=-\ell}^{\ell} a_{\ell m} Y_{\ell m}(\boldsymbol{n}) \tag{2.61}$$

where each $a_{\ell m}$ is the spherical harmonic coefficient. These coefficients can be calculated by integrating over the anisotropy data, i.e.,

$$a_{\ell m} = \int \Delta T(\boldsymbol{n}) Y^*_{\ell m} \, \mathrm{d}\boldsymbol{n} \,, \tag{2.62}$$

where $Y^*_{\ell m}$ indicates the complex conjugate of the spherical harmonics. The expectation value of each coefficient vanishes, i.e.,

$$\langle a_{\ell m} \rangle = 0 \,, \tag{2.63}$$

and the power spectrum C_ℓ is defined by

$$\langle a_{\ell m} a^*_{\ell' m'} \rangle = \delta_{\ell \ell'} \delta_{m m'} C_\ell \,, \tag{2.64}$$

where $a^*_{\ell' m'}$ indicates the complex conjugate of the spherical harmonic coefficient, and δ is the Kronecker delta,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases} \tag{2.65}$$

Figure 2.3 shows this power spectrum of CMB temperature anisotropies as detected by Planck (Planck Collaboration et al., 2018a). The red curve shows the theoretical

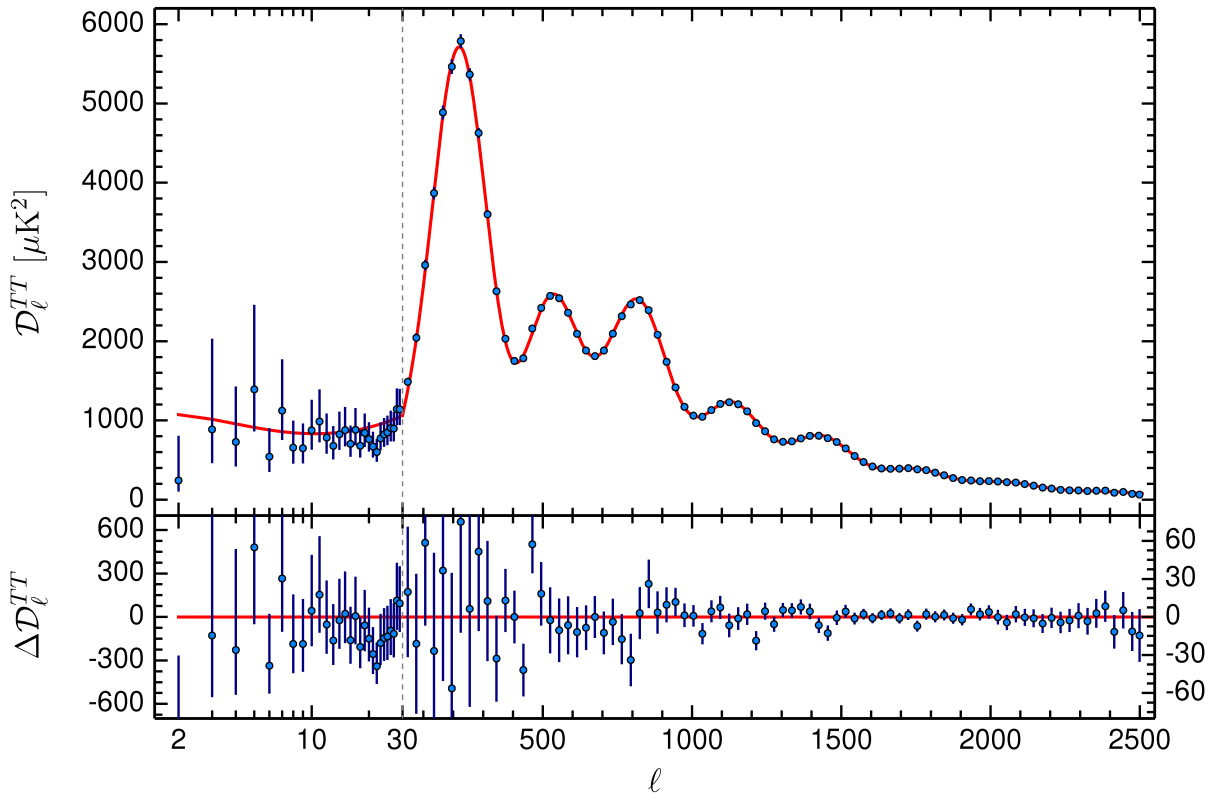

Figure 2.3: The cosmic microwave background temperature power spectrum measured by the Planck satellite (Planck Collaboration et al., 2016). The red curve shows the best fitting model, and the blue error bars show the measurements from Planck. The top panel shows the power spectrum, while the bottom panel shows the residuals. *Figure reproduced with permission from Planck Collaboration et al. (2016), copyright ESO.*

power spectrum predicted by the inferred cosmological model, which displays very good agreement with the observations.

2.2.2 Type-Ia supernovae

Supernovae are an explosive phase of a star's evolution. While there are many different types of supernovae, varying in the mechanism of their explosion and the progenitor star from which it was formed, type-Ia supernovae are of particular interest for cosmology. These supernovae are formed from white dwarfs, stellar remnants of low mass stars at the end of their life after fusion has stopped. White dwarfs are supported by electron degeneracy pressure. However, this pressure is only able to support the star when its

mass is $< 1.4M_{\odot}$, known as the Chandrasekhar mass (Chandrasekhar, 1931). If the white dwarf accretes matter from a nearby companion, this mass limit can be exceeded. Carbon fusion then begins in the core of the star, resulting in a thermonuclear runaway reaction (e.g., Nomoto, 1982); this is the supernova explosion.

Due to the standard explosion mechanism shared by all type-Ia supernovae, it might be hoped that the explosions were all of the same absolute magnitude. These supernovae could then be used as standard candles, objects with a known intrinsic brightness. Instead, the peak absolute magnitude of these supernovae actually varies significantly (e.g., Phillips et al., 1999). However, the intrinsic magnitude of supernovae also varies in time, allowing the plotting of a *light curve* of this magnitude against time. It was discovered by Phillips (1993) that the peak absolute magnitude of these supernovae are correlated with the rate at which this magnitude decreases. Specifically, the brighter the peak magnitude of the supernova, the longer this decrease in magnitude takes. Thus, by observing a supernova over the several days this decrease takes and constructing a light curve, it is possible to correct the absolute magnitude (e.g., Hamuy et al., 1996). The resulting corrected magnitude than has a scatter of $\approx 10^{-1}$ mag between supernovae. Type-Ia supernovae are therefore said to be *standardisable candles*. Modern approaches to supernova cosmology utilise sophisticated light curve fitting methods such as SALT2 (Guy et al., 2007) to make these corrections.

By using these correction methods, the peak absolute magnitude of the supernova M is known, while the apparent magnitude m is observed. These quantities can then be used to define the *distance modulus* as

$$\mu = m - M \,. \tag{2.66}$$

This distance modulus is related to the luminosity distance D_L by (Hogg, 1999)

$$\mu = 5 \log_{10} \left(\frac{D_L}{10 \text{ pc}} \right) , \tag{2.67}$$

where pc refers to *parsec*, an astronomical unit of distance[9]. Finally, the luminosity distance is related to the redshift of the supernova z by equation 2.47. Spectroscopic observations of the host galaxy of the supernova can be used to obtain this redshift, though photometric redshifts can be obtained from the light curve fit (e.g., Palanque-Delabrouille et al., 2010) if this is not possible[10].

[9]The parsec is defined as the distance at which an object observed from Earth would have a parallax of one arcsecond, hence the name.

[10]Photometric redshift methods used for supernovae rely on fitting the light curve, i.e., observations of the supernova over time. These methods are therefore distinct from the photometric redshift

Since the distance-redshift relation is cosmology dependent, observations of the distance moduli of supernovae can be used to constrain cosmological parameters. Observations made in this way by Riess et al. (1998) and Perlmutter et al. (1999) gave the first evidence that the expansion of the universe was accelerating as discussed in section 2.1.8.

2.2.3 Baryon acoustic oscillations

Before recombination as described in section 2.2.1, baryonic matter was coupled to photons. Thus, while dark matter clustered freely, baryonic matter experienced a radiation pressure, causing the propagation of sound waves. After recombination, baryonic matter is preferentially found in shells surrounding overdensities of dark matter, the size of which is dependent on cosmology.

These *baryon acoustic oscillations* (BAOs) can now be observed in the distribution of galaxies as a preferred distance of separation. As this physical size is set by cosmology, these BAOs are an observational feature with a known, fixed size. This is analogous to standard candles, sources with a known luminosity. Thus, BAOs are an example of a *standard ruler*.

Observations of BAOs in the distribution of galaxies puts constraints on cosmological parameters. By measuring their angular size on the sky, their angular diameter distance D_A described in section 2.1.7 can be calculated. This distance depends both on cosmological parameters and the redshift z. Thus, by inferring the redshift to the observed galaxies through either spectroscopy or photometric redshifts as discussed in chapter 4, cosmological parameters can be inferred.

BAOs were first detected in the correlation function of galaxies from the Sloan Digital Sky Survey (SDSS) by Eisenstein et al. (2005). These observations indicated that the size of these oscillations was ≈ 150 Mpc at the time of recombination, and the cosmology inferred was consistent with ΛCDM.

2.3 Cosmology with Photometric Galaxy Surveys

The photometric redshift methods discussed throughout the research work of this thesis are designed for use with cosmological galaxy surveys. This section introduces these surveys and discusses how they are used to constrain cosmology.

methods discussed throughout this thesis which do not rely on observing light curves and are therefore suitable for inferring redshifts of non-transient objects such as galaxies.

2.3.1 Spectroscopic and photometric galaxy surveys

Before we discuss how galaxy surveys can be used to constrain cosmological parameters, we take an aside to describe the two types of galaxy survey; spectroscopic and photometric. These two types of survey differ in their method of observation, and it is the latter we are concerned with throughout this thesis.

When observing a galaxy in a cosmological galaxy survey, we are interested not only in its angular position on the sky, but also in its redshift. This provides three-dimensional information about its position, allowing statistical statements to be made about the large-scale distribution of galaxies in the late-time universe. Spectroscopic and photometric galaxy surveys then differ in the method they use to obtain this redshift.

Spectroscopic redshifts obtain the redshift of a galaxy by observing its spectrum, the flux of the galaxy observed at specific wavelengths. To obtain this spectrum, the light collected by the telescope is dispersed before being focussed onto the detector. This dispersion is typically accomplished using a diffraction grating or a *grism*, the combination of a prism and a grating (e.g., Greene et al., 2016). This spectrum is then used to identify emission and absorption lines caused by specific elements present within the stars and the surrounding interstellar medium comprising the galaxy. These emission and absorption lines occur at fixed, specific wavelengths in the rest frame. Thus, the redshift of the galaxy can be inferred from the spectrum by identifying a shift in the wavelengths of these lines, e.g., by cross-correlating the observed spectrum with a bank of template spectra (e.g., Tonry and Davis, 1979; Baldry et al., 2014) or fitting the lines directly with Gaussian profiles (e.g., Mink and Wyatt, 1995).

Photometric redshifts are a statistical method of inferring the redshift of a galaxy from broadband photometry, typically a small number of fluxes measured from images obtained using colour filters. These filters allow only a limited range of wavelengths observed by the telescope to be detected, e.g., the five optical *ugriz* filters of SDSS have a full-width-half-maximum (FWHM)[11] of $\approx 600 - 1500 \times 10^{-1}$ m (Fukugita et al., 1996). This photometry therefore acts as an extremely low-resolution spectrum of the galaxy. Photometric redshifts are the subject of the research work of this thesis, and are discussed in more detail in chapter 4

The advantage to inferring redshifts spectroscopically is their precision. Modern spectroscopic redshift approaches can identify the recessional velocity of galaxies to a precision of ≈ 50 km s^{-1} (Baldry et al., 2014), corresponding to a negligible uncertainty

[11] The descriptively named FWHM of a curve is its complete width, measured at half of its maximum height. For a Gaussian of standard deviation σ, the FWHM $\approx 2.36\sigma$.

in the redshift for cosmological applications. However, the cost of this precision is that spectroscopic redshifts are expensive in terms of telescope time. Dispersing the the observed light in order to obtain a spectrum reduces the signal-to-noise of the detection, necessitating long integration times. Since the colour filters used when obtaining images for photometric redshifts are substantially wider than the resolution of a spectrograph, the signal-to-noise of the detection is greater for a fixed integration time. As a result, spectroscopic redshifts cannot be obtained for sources as faint as when using photometric redshifts.

In addition, spectroscopy cannot be used to obtain redshifts for large numbers of galaxies simultaneously, with fibre-based spectrographs able to observe $\approx 10^2$ objects in a single telescope pointing (e.g., Kimura et al., 2010). In contrast, photometric redshifts can be measured for all galaxies detected in an image, and are therefore limited in the number of galaxies observed simultaneously only by the field of view and resolution of the telescope. As a result, photometric redshifts can be used to infer redshifts of samples of galaxies that are greater in both number and depth than can be achieved with spectroscopic redshifts, at the cost of a reduction in precision.

2.3.2 The matter power spectrum

The observable statistical quantities associated with photometric galaxy surveys are related to cosmology through the matter power spectrum $P(k)$. This section briefly discusses this quantity, its interpretation, and how it is related to cosmological parameters.

The statistical distribution of matter comprising the large-scale structure of the universe is described in terms of the density contrast $\delta(\boldsymbol{x})$, the fractional overdensity of matter at $\boldsymbol{x}$, defined by

$$\delta(\boldsymbol{x}) = \frac{\rho(\boldsymbol{x}) - \bar{\rho}}{\bar{\rho}}\,, \tag{2.68}$$

where $\rho(\boldsymbol{x})$ is the density of matter at $\boldsymbol{x}$, and $\bar{\rho} \equiv \rho_c \Omega_m$ is the average matter density. We can also define the density contrast field in the frequency domain by taking its Fourier transform like

$$\delta(\boldsymbol{k}) = \int \delta(\boldsymbol{x}) \mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{x}}\, \mathrm{d}^2 x\,. \tag{2.69}$$

Since the density contrast is a real field, the Fourier components obey a Hermitian symmetry

$$\delta(\boldsymbol{k}) = \delta^*(-\boldsymbol{k})\,, \tag{2.70}$$

where δ^* denotes the complex conjugate of δ.

This density contrast is a random field, a stochastic object described by a joint probability distribution (e.g., Bardeen et al., 1986), e.g.,

$$P(\delta(\boldsymbol{x}_1), \delta(\boldsymbol{x}_2) \dots \delta(\boldsymbol{x}_N))\ \mathrm{d}\delta(\boldsymbol{x}_1)\ \mathrm{d}\delta(\boldsymbol{x}_2) \dots\ \mathrm{d}\delta(\boldsymbol{x}_N)\,. \tag{2.71}$$

As described in section 2.1.2, it is generally assumed that the universe obeys the cosmological principle; that is, that the universe is homogeneous and isotropic on sufficiently large scales. These properties can be understood in terms of this random field as follows. Firstly, the field is said to be homogeneous if these joint probability distribution functions are invariant under translation $\boldsymbol{x} \to \boldsymbol{x} + \boldsymbol{r}$ for an arbitrary vector $\boldsymbol{r}$. In addition, the field is said to be isotropic if the joint probability distribution functions are invariant under rotation $\boldsymbol{x} \to \mathcal{R}\boldsymbol{x}$ for an arbitrary rotation matrix $\mathcal{R}$.

A particularly convenient type of random field is a *Gaussian* random field. In this case, the coefficients of the Fourier series expansion of the field are independent from each other and have phases which are random (Bardeen et al., 1986). Such a field is statistically described entirely by its two-point statistics; all higher-order correlations are zero. On sufficiently large scales, a Gaussian random field will remain Gaussian after gravitational evolution. Inflationary models predict that the quantum fluctuations of the early universe will produce a Gaussian initial density field (Bardeen et al., 1983). Thus, we expect that the large-scale matter distribution can be described as a Gaussian random field.

The propensity for the field to cluster can be quantified statistically through the correlation function $\xi(r)$. More specifically, it is a measure of the excess probability of finding two particles separated by a distance $r_{12} = |\boldsymbol{r}_1 - \boldsymbol{r}_2|$ within volume elements $\mathrm{d}V_1$ and $\mathrm{d}V_2$ over random probability. This can be defined as

$$\mathrm{d}P_{12} = n^2\,[1 + \xi(r_{12})]\ \mathrm{d}V_1\ \mathrm{d}V_2\,, \tag{2.72}$$

where n is the mean density of particles. More generally, this can be extended to higher order statistics, e.g., for the three point correlation function

$$\mathrm{d}P_{123} = n^2\,[1 + \xi(r_{12}) + \xi(r_{23}) + \xi(r_{31}) + \xi_3(r_{12}, r_{23}, r_{31})]\ \mathrm{d}V_1\ \mathrm{d}V_2\ \mathrm{d}V_3\,. \tag{2.73}$$

However, for Gaussian fields, $\xi_3(r_{12}, r_{23}, r_{31}) = 0$, since the field is fully specified through the two point statistics. This probabilistic definition of the correlation function gives a physical interpretation, but it can also be defined as an ensemble average over the density contrast field by

$$\xi(r) = \langle \delta(\boldsymbol{x})\delta(\boldsymbol{x} + \boldsymbol{r}) \rangle\ , \tag{2.74}$$

where $r \equiv |\boldsymbol{r}|$ due to isotropy.

Analogously to this correlation function, we can define the *matter power spectrum* to be an ensemble average over the density field defined in Fourier space, giving

$$\begin{aligned}\langle \delta(\boldsymbol{k})\delta^*(\boldsymbol{k}')\rangle &= (2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')\int \mathrm{d}^3\boldsymbol{r}\xi(r)\mathrm{e}^{-i\boldsymbol{k}\cdot\boldsymbol{r}} \\ &\equiv (2\pi)^3\delta_D(\boldsymbol{k}-\boldsymbol{k}')P(k)\,,\end{aligned} \tag{2.75}$$

where $\delta_D(\boldsymbol{x})$ is the Dirac delta function. The correlation function and the matter power spectrum are therefore defined to be a Fourier transform pair, given by

$$P(k) = \int \mathrm{d}^3 r\xi(r)\mathrm{e}^{-i\boldsymbol{r}\cdot\boldsymbol{k}} \tag{2.76}$$

and

$$\xi(r) = \int \frac{\mathrm{d}^3k}{(2\pi)^3}P(k)\mathrm{e}^{i\boldsymbol{r}\cdot\boldsymbol{k}}\,. \tag{2.77}$$

Like the definition for the correlation function, the power spectrum depends only on $k = |\boldsymbol{k}|$ due to the isotropy of the density field. In general, since the statistical distribution of matter evolves through gravitational interaction, the matter power spectrum also has a redshift dependence $P(k, z)$.

The typical method to predict the matter power spectrum for a given cosmology is the halo model (Peacock and Smith, 2000; Seljak, 2000). This is an analytic model describing the cosmological matter distribution as a series of spherically symmetric haloes, the spatial distribution of which is clustered. By constraining the free parameters of the model using N-body simulations, the matter power spectrum can be predicted over a range of cosmologies (e.g., Smith et al., 2003). An alternative approach is to construct *emulators* which interpolate between matter power spectra obtained from N-body simulations using machine learning methods such as Gaussian processes (e.g., Heitmann et al., 2016).

Effects of baryonic feedback on the matter power spectrum

Since the matter content of the Universe is largely dominated by dark matter (Planck Collaboration et al., 2018a), many N-body simulations have neglected the effect of baryons entirely, modelling only dark matter particles. Despite this however, the effect of these baryons on the matter power spectrum remains significant and their exclusion can result in large systematic errors in power spectrum models from these simulations. These effects are brought about due to baryonic process including star formation, radiative cooling, and feedback from active galactic nuclei (AGN) and supernovae (Rudd

et al., 2008).

The effects of baryons on the matter power spectrum are greatest at small scales (Harnois-Déraps et al., 2015). Before the formation of structure, the distribution of baryons traced that of dark matter, However, unlike dark matter which does not interact electromagnetically, baryons are able to dissipate energy through radiation. This causes them to condense into the centres of dark matter halos at greater densities (Rudd et al., 2008), increasing small-scale power in the matter power spectrum.

AGN are extremely high luminosity astrophysical sources at the centre of galaxies (Osterbrock, 1991). Their luminosity is driven by accretion of matter from a disc onto a central black hole. In addition, some AGN produce relativistic jets, rapid outflows of matter in highly collimated beams that lie along the axis of rotation of the disc. AGN are fuelled by the matter which surrounds them, but the large pressure from jets and radiation pushes matter away from the black hole. This causes the luminosity of the AGN, and thus this pressure, to fall, until matter is able to condense again and the pressure returns. This effect is known as AGN feedback (Fabian, 2012).

The result of this AGN feedback is that matter is transferred into lower density regions from high density regions (Rudd et al., 2008), causing a suppression in the matter power spectrum at small scales. However, if this feedback only affects the nearby environment, clustering of AGNs causes these perturbations at small scales to be pushed to larger scales; this is a competing effect. The interactions between the AGN and the surrounding matter are therefore complicated, making them difficult to model.

In order to not bias inference of cosmological parameters, the effect of baryonic feedback should be included in predictions of the matter power spectrum. By comparing the statistics of the matter distribution in N-body simulations with and without baryonic feedback, these effects have been found to cause a suppression in the $z = 0$ matter power spectrum of $\approx 10 - 20\%$ at small scales $k \approx 10$ h Mpc^{-1} (Chisari et al., 2018; Schneider et al., 2019). Methods for predicting $P(k, z)$ that can account for these effects have therefore been developed (e.g., Harnois-Déraps et al., 2015; Mead et al., 2015).

2.3.3 Constraining cosmology with 3 × 2 pt. analyses

Cosmological parameters are constrained using photometric galaxy surveys by measuring angular power spectra from observations. These angular power spectra can also be computed from theory through the matter power spectrum described in section 2.3.2,

allowing these observations to be used to make inferences about cosmology. A common approach to this is to use a 3×2 pt. analysis; that is, constraining cosmology through observation of three distinct sets of two-point statistics. These are described below.

Photometric galaxy surveys image large populations of galaxies. Galaxies are separated into several redshift bins, a process known as *tomography* (e.g., Hu, 1999; Petri et al., 2016; Joudaki et al., 2018). This allows galaxy surveys to probe large-scale structure as a function of redshift without using a full three-dimensional analysis (Heavens, 2003). A tomographic analysis is also more well-suited to using low-precision photometric redshifts.

For some measurements such as galaxy-galaxy lensing, these galaxy populations are also separated into two distinct sets; lens galaxies and source galaxies. Lens galaxies are lower redshift than source galaxies, and observations of these galaxies are restricted to their position. In contrast, source galaxies also have their *shape* measured. When the light from these galaxies is perturbed through gravitational lensing, the ellipticities of nearby source galaxies become correlated[12]. Thus, by measuring the ellipticities of many galaxies near each other on the sky, the effect of lensing can be constrained. Since this lensing depends on the distribution of matter along the line of sight to the galaxy, measuring galaxy ellipticities provides a direct probe of matter in the universe.

The tomographic bins of each of the sets of galaxies can then be correlated in the three following ways to construct the 3×2 pt. spectra. Firstly, the positions of galaxies in tomographic bin i can be correlated with the positions of galaxies in tomographic bin j, where i and j can be either different or equal; this is known as *galaxy clustering*. Secondly, the positions of lens galaxies in a tomographic bin can be correlated with the shapes of source galaxies in another tomographic bin; this is known as *galaxy-galaxy lensing*. Finally, the shapes of source galaxies in tomographic bin i can be correlated with the shapes of source galaxies in tomographic bin j, where i and j can again be either different or equal; this is known as *cosmic shear*.

We now review the calculation of the 3×2 pt. spectra, following Krause et al. (2017). The power spectra correlating tomographic bins i and j can be calculated using the limber approximation (Limber, 1953) for galaxy clustering as

$$C_{gg}^{ij}(\ell) = \int \frac{q_g^i\left(\frac{\ell+0.5}{\chi},\chi\right) q_g^j\left(\frac{\ell+0.5}{\chi},\chi\right)}{\chi^2} P\left(\frac{\ell+0.5}{\chi}, z(\chi)\right)\, \mathrm{d}\chi\,, \tag{2.78}$$

[12]In practice, the ellipticities of nearby galaxies can be correlated without lensing, an effect known as *intrinsic alignments*. See the reviews of Joachimi et al. (2015) and Troxel and Ishak (2015) for more details.

for galaxy-galaxy lensing as

$$C^{ij}_{g\kappa}(\ell) = \int \frac{q^i_g\left(\frac{\ell+0.5}{\chi}, \chi\right) q^j_\kappa(\chi)}{\chi^2} P\left(\frac{\ell+0.5}{\chi}, z(\chi)\right) \, \mathrm{d}\chi \,, \tag{2.79}$$

and for cosmic shear as

$$C^{ij}_{\kappa\kappa}(\ell) = \int \frac{q^i_\kappa(\chi)\, q^j_\kappa(\chi)}{\chi^2} P\left(\frac{\ell+0.5}{\chi}, z(\chi)\right) \, \mathrm{d}\chi \,, \tag{2.80}$$

where $P(k, z)$ is the matter power spectrum described in section 2.3.2, χ is the comoving distance, and $z(\chi)$ is the redshift corresponding the comoving distance χ, given by the inverse of equation 2.43. These equations also involve integrations over the radial weight function q_g and the lensing efficiency q_κ. Assuming a linear galaxy bias b^i, these are given by

$$q_g(k, \chi) = b^i \frac{n^i(z(\chi))}{\bar{n}^i} \frac{\mathrm{d}z}{\mathrm{d}\chi} \,, \tag{2.81}$$

and

$$q_\kappa(k, \chi) = \frac{3H_0^2\Omega_m}{2c^2} \frac{\chi}{a(\chi)} \int_\chi^{\chi_H} \frac{n^i(z(\chi'))}{\bar{n}^i} \frac{\mathrm{d}z}{\mathrm{d}\chi'} \frac{\chi' - \chi}{\chi'} \, \mathrm{d}\chi' \,, \tag{2.82}$$

where χ_H is the comoving horizon[13], $a(\chi)$ is the scale factor at comoving distance χ, $n^i(z)$ is the redshift distribution of galaxies in tomographic bin i, and $\bar{n}^i$ is the number density of galaxies in this bin, calculated by integrating the distribution over redshift as

$$\bar{n}^i = \int n^i(z) \, \mathrm{d}z \,. \tag{2.83}$$

Finally, these power spectra can be transformed to angular correlation functions that can be measured from the observed galaxies. For galaxy clustering, this is given by

$$w^i(\theta) = \sum_\ell \frac{2\ell+1}{4\pi} P_\ell\left(\cos(\theta)\right) C^{ii}_{gg}(\ell) \,, \tag{2.84}$$

where $P_\ell(\dots)$ is a Legendre polynomial of order ℓ, for galaxy-galaxy lensing by

$$\gamma^{ij}(\theta) = \int \frac{\ell}{2\pi} J_2(\ell\theta) C^{ij}_{g\kappa}(\ell) \, \mathrm{d}\ell \,, \tag{2.85}$$

and for cosmic shear by

$$\xi^{ij}_\pm(\theta) = \int \frac{\ell}{2\pi} J_{0/4}(\ell\theta) C^{ij}_{\kappa\kappa}(\ell) \, \mathrm{d}\ell \,, \tag{2.86}$$

[13]The comoving horizon is the maximum comoving distance a photon could have propagated, given the age of the universe. This can be calculated using equation 2.42, where $t_e = 0$ and t_o is the current age of the universe.

where $J_x(\dots)$ is a Bessel function of order x.

By using these equations to calculate theoretical correlation functions from cosmology through the matter power spectrum, observations from photometric galaxy surveys can constrain cosmological parameters. The above process demonstrates the two distinct needs for photometric redshifts of galaxies. Firstly, these redshifts are required to place each galaxy into its corresponding tomographic bin. Secondly, evaluating the weighting functions q_g and q_κ requires the redshift distributions of galaxies in each tomographic bin. As discussed in section 2.3.1, galaxy surveys that constrain cosmology in this way cannot utilise spectroscopic redshifts due to the number and depth of galaxies observed, necessitating photometric redshifts. The applications of photometric redshifts to each of these distinct uses are discussed in chapter 4.

2.3.4 Tensions and open problems

While ΛCDM has survived many observational tests, some open problems still remain. During this section, we briefly discuss some of these challenges that can be addressed using photometric galaxy surveys.

Hubble constant tension

As described in section 2.1.1, the Hubble constant H_0 describes the current rate of expansion of the universe. The value of this constant can be inferred from several different cosmological probes. Firstly, observations of the CMB can constrain its value by assuming a cosmological model, as detailed in section 2.2.1. Observations from Planck (Planck Collaboration et al., 2018a) have inferred its value to be $H_0 = (67.36 \pm 0.54)$ km s^{-1} Mpc^{-1}.

In addition, it is also possible to make a *local* measurement of the Hubble constant. This uses a series of standard candles to construct the *distance ladder*, allowing the determination of the distance of cosmological objects and, by measuring their redshifts, infer H_0 through Hubble's law defined in equation 2.1. This measurement was made by Riess et al. (2016), finding the value $H_0 = (73.24 \pm 1.74)$ km s^{-1} Mpc^{-1}, significantly higher than the value from Planck.

The precision of these two mutually incompatible values means that these results are in tension. Since the CMB-derived value relies on the assumption of a cosmological model, it is possible that this discrepancy points to the existence of new physics. However, it is also possible that an unknown systematic error in one or both of these

measurements is responsible, and that accounting for this uncertainty would resolve the tension. It is therefore important to find independent probes that can also measure the Hubble constant and inform this.

Dark Energy Survey Collaboration et al. (2018) uses weak lensing and galaxy clustering measurements as described in section 2.3.3 to constrain cosmological parameters using observations from the Dark Energy Survey (Dark Energy Survey Collaboration et al., 2016). By combining these constraints with those from BAO (Beutler et al., 2011; Ross et al., 2015; Alam et al., 2017) and Big Bang nucleosynthesis (BBN, Fixsen, 2009; Cooke et al., 2016) experiments, they find the value of the Hubble constant to be $67.4^{+1.1}_{-1.2}$ km s^{-1} Mpc^{-1}. This is an independent constraint which agrees with the value of H_0 determined by Planck, demonstrated in the contours shown in Figure 2.4. As future galaxy surveys bring increased precision, constrains made in this way from galaxy surveys could help to discriminate between in-tension measurements of the Hubble constant.

$\Omega_{m,0} - \sigma_8$ tension

Unlike the Hubble constant, the $\Omega_{m,0} - \sigma_8$ tension is a tension where the results from galaxy surveys disagree with those from Planck CMB measurements. As described in section 2.1.6, $\Omega_{m,0}$ is the present-day matter density parameter. The parameter σ_8 is the r.m.s. linear matter power spectrum fluctuations in spheres of radius 8 h^{-1} Mpc.

It is possible to infer both of these parameters from both photometric galaxy surveys and CMB observations. Figure 2.5 shows a comparison of contours for both of these values. This shows that the values obtained from Planck CMB observations (Planck Collaboration et al., 2016) are in tension with those obtained from the KiDS galaxy survey (Hildebrandt et al., 2017). A future high-precision galaxy survey could clarify this tension and determine whether it is the result of a systematic or statistical uncertainty, or the result of new physics.

Constraining the sum of the neutrino masses

Neutrinos are low mass particles first postulated by Pauli (1930) to preserve conservation of energy and momentum during beta decay, and are now described by the standard model of particle physics. Neutrinos interact through the weak force, mediated by W^+, W^- and Z^0 bosons (Dolgov, 2002), though are electromagnetically neutral. The standard model predicts three active species of neutrino, electron (ν_e), muon (ν_μ) and tau (ν_τ) Ramond (1999). This is supported by observations which

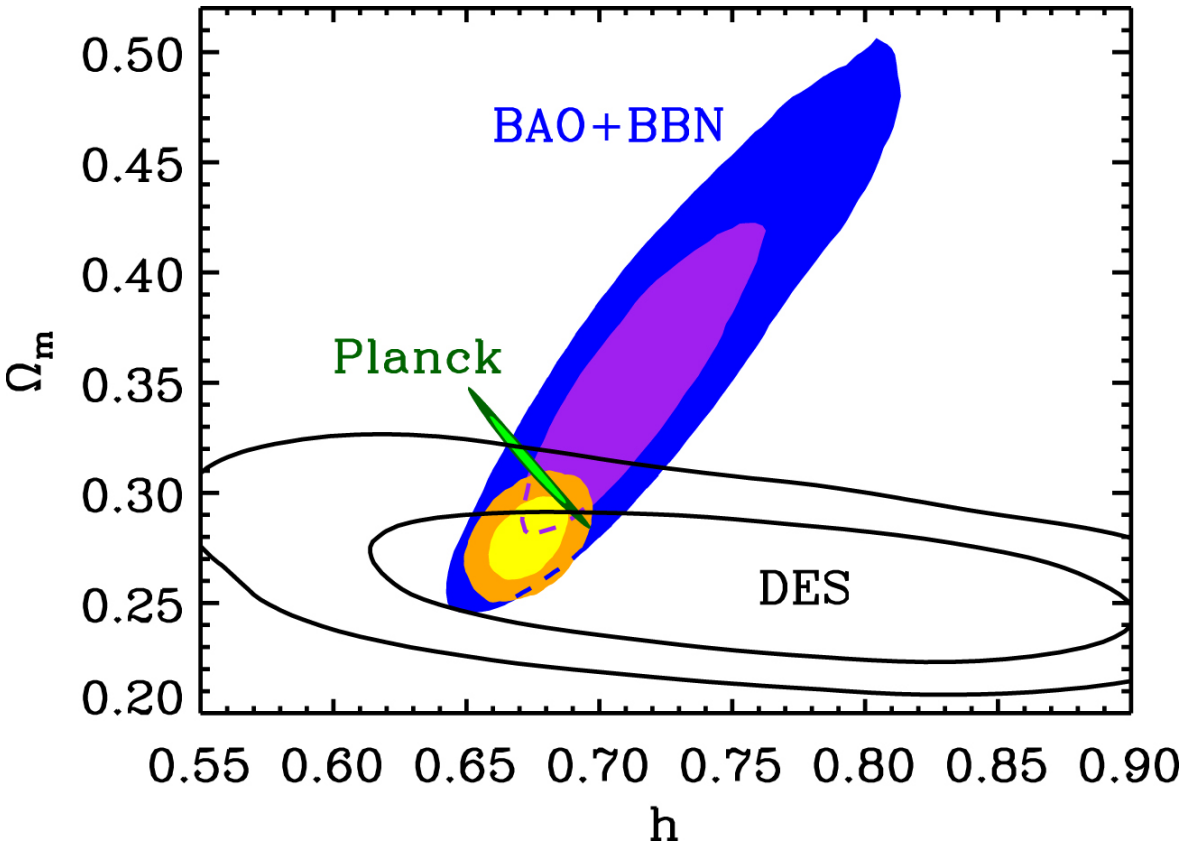

Figure 2.4: Contours showing joint $\Omega_m - h$ constraints for a variety of cosmological probes, where the Hubble constant $H_0 \equiv 100\,h$ km s^{-1} Mpc^{-1}. The blue contours show the results for BAO (Beutler et al., 2011; Ross et al., 2015; Alam et al., 2017) and BBN (Fixsen, 2009; Cooke et al., 2016), the unfilled black contours show the results from the Dark Energy Survey (Dark Energy Survey Collaboration et al., 2018) alone, and the yellow contours show the results from the combination of these datasets. These are compared to the green contours, showing the results from Planck observations of the CMB (Planck Collaboration et al., 2018a). *Figure taken with permission from Figure 1 of Dark Energy Survey Collaboration et al. (2018).*

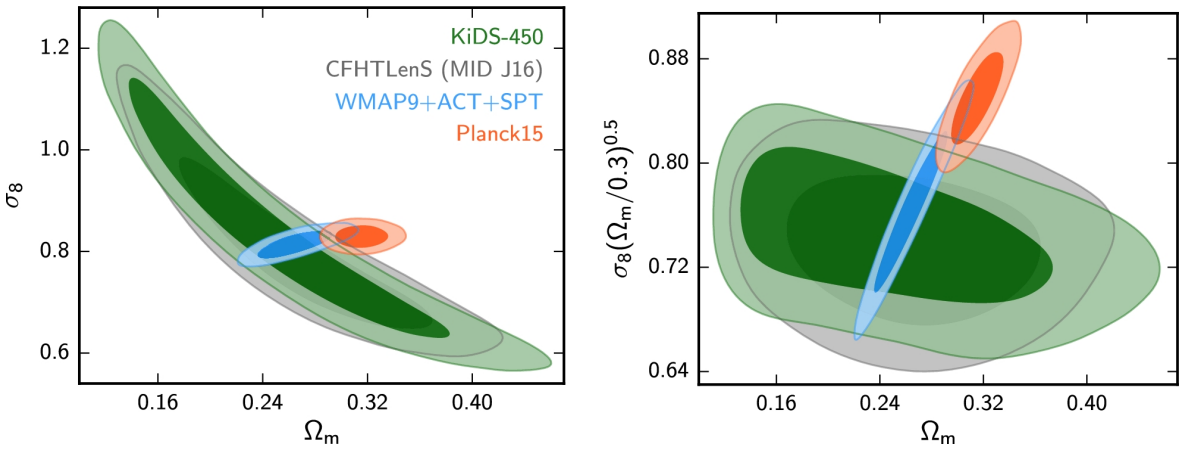

Figure 2.5: A comparison of posterior contours for the $\Omega_{m,0}$ and σ_8 parameters. The left panel shows the contours of these parameters directly, while the right panel shows the reparametrisation $S_8 \equiv \sigma_8 \sqrt{\Omega_m/0.3}$. The red contours show the results from Planck CMB observations (Planck Collaboration et al., 2016), while the green contours show results obtained from the KiDS galaxy survey (Hildebrandt et al., 2017). The contours show that these two results are in tension. *Figure taken with permission from Figure 6 of Hildebrandt et al. (2017).*

demonstrate an effect known as *neutrino oscillation*, whereby particles of one species can convert to another species spontaneously King (2007).

Although neutrino oscillations support the prediction of the standard model that three species of neutrino exist, they add an additional conflict. While the standard model predicts strictly massless neutrinos, these oscillations can only occur due a difference in mass between the flavours; thus, at least two of these neutrino flavours must be massive. These observations also allow the possiblilty of two secenarios, known as the *normal* and *inverted* hierarchies, depending on the ordered of the flavour masses (e.g., Qian and Vogel, 2015). This conflict therefore presents evidence of beyond the standard model (BSM) physics, an exciting prospect for an otherwise extraordinarily successful theory.

These neutrino oscillation observations can place lower limits on the sum of the neutrino masses, labelled $\sum m_\nu$. These indicate that $\sum m_\nu > 60$ meV assuming a normal hierarchy, and $\sum m_\nu > 100$ meV for an inverted hierarchy. Cosmology is also able to place an upper limit on their mass, as the free-streaming of neutrinos from high-density regions causes a mass-dependent suppression in the matter power spectrum (e.g., Allison et al., 2015). By combining a variety of cosmological probes, Palanque-Delabrouille et al. (2015) found an upper limit of $\sum m_\nu < 120$ meV, while Choudhury and Choubey (2018) combine CMB temperature and polarisation, BAO and supernovae data to obtain an upper limit of $\sum m_\nu < 118$ meV. However, Loureiro et al. (2019) caution that the choice of neutrino model can substantially alter the

resulting bound; they report an upper limit of $\sum m_\nu < 260$ meV when assuming a model that is consistent with particle physics experiments, substantially higher than the upper limit of $\sum m_\nu < 150$ meV found by neglecting this information.

Observations from future high-precision galaxy surveys could increase the strength of this bound, potentially determining the nature of the mass hierarchy by excluding the lower limit of the inverted hierarchy.

2.3.5 Future photometric galaxy surveys

Of particular interest for the research work of this thesis is the future photometric galaxy survey known as the Large Synoptic Survey Telescope (LSST, Ivezić et al., 2019). This will use an 8.4 m mirror to survey the southern sky in optical wavelengths to a depth of $m_r < 27$ (LSST Science Collaboration et al., 2009). Among a variety of other science goals, this will allow high-precision cosmological constraints to be made using the methods described in section 2.3.3. However, since LSST is a ground based telescope, this high depth of photometry will result in around half of all galaxies observed being blended (Dawson and Schneider, 2014), i.e., overlap with other galaxies along the line of sight; this problem of blending is discussed further in chapter 5.

Another future photometric galaxy survey is Euclid (Laureijs et al., 2011). This will be a 1.2 m space-based near-infrared observatory that will measure weak lensing and BAO in order to constrain cosmology. Due to the smaller design than LSST, owing to it being space-based rather than ground-based, Euclid will be less sensitive, reaching a depth of 24 mag over three near-infrared bands and 24.5 mag in a wider optical band. However, the advantage of a space-based design is in resolution, as Euclid will not be hampered by the atmospheric effects that impact ground-based telescopes such as LSST. As a result, some sufficiently-bright sources that are blended in LSST will be resolved in Euclid, potentially aiding the deblending process (Rhodes et al., 2017). This scenario, which we refer to as *partial-blending*, is discussed in more detail in chapter 6.

Chapter 3

Statistical Methodology

The use of statistical analysis is a fundamental part of the scientific method, the central tenet of which is empiricism; a scientist uses observations of the world in order to draw conclusions and update their beliefs. Reasoning about phenomena in this way is inevitably a probabilistic exercise. Observations and experiments are never perfect in the sense that they provide a definitive measurement, but rather are subject to stochastic errors. These statistical, or *aleatoric*, uncertainties limit how informative the data can be about a particular phenomena. The models and assumptions used to analyse these data are also never perfect in practice, resulting in systematic, or *epistemic*, uncertainties. These various sources of uncertainties must be incorporated into the analysis if the conclusions are to be an accurate reflection of the state of our knowledge. This is the role of statistical inference.

Recent years have seen cosmology transition from a data-poor science into one with an abundance of data. As both the volume and quality of data increase, statistical uncertainties are reduced, giving rise to the era of precision cosmology. It is therefore increasingly important that statistical analyses are rigorous in their accounting of uncertainties, since the increased significance of any resulting biases could mistakenly point to the existence of new physics. With several upcoming experiments such as the Large Synoptic Survey Telescope (LSST, Ivezić et al., 2019) and Square Kilometre Array (SKA, Dewdney et al., 2009) becoming available, this trend of increased volume of data is set to continue. This *big data* scenario presents a significant computational challenge for statistical inference, as methods must scale efficiently to these very large datasets. The higher precision constraints afforded by these increased sizes of datasets also mean that systematic effects that were previously neglected must now be accounted for.

This chapter introduces and discusses several statistical methodologies that are

used throughout this thesis. Section 3.1 introduces Bayesian statistics, a natural method for making inferences about scientific problems from observations while accounting for uncertainties. Section 3.2 discusses machine learning, a series of methods that utilise flexible models to make predictions informed only by previously seen training data, rather than a specific physical model.

3.1 Bayesian Inference

Much of the work in the thesis makes use of Bayesian inference methods. This section introduces these methods, discusses the problems they can be applied to and the considerations that must be made to apply them.

3.1.1 Bayesian and frequentist interpretations of probability

Before detailing how probabilities can be computed and manipulated mathematically, it is worth taking an aside to discuss the interpretation of probability — what does it mean to ascribe a probability to a particular proposition, and what are we able to conclude as a result? There are two contrasting views on this question; probabilities may be given either a *frequentist* or *Bayesian* interpretation (e.g., Bayarri and Berger, 2004). Here, we describe their differences and argue for the use of Bayesian methods for cosmological data analysis.

As implied by the name, a frequentist interpretation views probability as the frequency of an event in a sequence of repeated trials, in the limit where the number of trials tends to infinity. A prototypical example of such an interpretation is that of flipping a coin. Under a frequentist interpretation, the probability of the coin landing on tails is defined as the fraction of tosses that landed on tails in an infinite sequence of tosses, i.e.,

$$P(\text{tails}) = \lim_{n_{\text{trials}} \to \infty} \frac{n_{\text{tails}}}{n_{\text{trials}}} \,. \tag{3.1}$$

Conclusions about the experiment can then be made using the likelihood $\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{d}) \equiv P(\boldsymbol{d} \mid \boldsymbol{\theta})$, a function that specifies the probability of a set of data $\boldsymbol{d}$, given the parameters $\boldsymbol{\theta}$. In this coin flipping example, a frequentist analysis can ask how probable a particular sequence of heads and tails is, assuming that the coin were fair, i.e., $P(\text{heads}) = P(\text{tails}) = 0.5$. It should be noted that the form of this question is different from that of interest in most scientific enquiries; inferring that a dataset would be unlikely under a particular hypothesis is not the same as inferring how likely that hypothesis is.

In practice, no dataset is infinite in size, meaning that the interpretation of frequentist methods must instead rely on asymptotic properties. Consider the example of a maximum-likelihood estimate (MLE). Assume we have a dataset $\{\boldsymbol{d}\} = \{\boldsymbol{d}_1, \boldsymbol{d}_2 \dots\}$ comprising many independent and identically distributed (i.i.d.) samples , drawn from a distribution dependent on a true parameter $\boldsymbol{\theta}_{\text{true}}$, i.e.,

$$\boldsymbol{d}_i \sim P(\boldsymbol{d} \mid \boldsymbol{\theta}_{\text{true}})\,. \tag{3.2}$$

The MLE $\hat{\boldsymbol{\theta}}$ is the value of the parameter that maximises the likelihood. However, this value depends on the particular dataset which has been randomly sampled. Different datasets will therefore result in different values of $\hat{\boldsymbol{\theta}}$, meaning that the estimate itself is a random variable, the distribution of which is known as the sampling distribution of this particular statistic. In the limit of the number of samples tending to infinity, this sampling distribution is asymptotically normal (e.g., Efron, 1982).

These asymptotic properties allow the result of frequentist inference to be formulated as a *confidence interval* corresponding to a particular percentage confidence. Given many repetitions of the experiment, each resulting in a different confidence interval, the corresponding percentage of these intervals should contain the true parameter value $\boldsymbol{\theta}_{\text{true}}$. It is once again worth noting the distinction between this definition and the question of what the probability of the true parameter being contained within a particular interval is. A frequentist interpretation of probability views population parameters as fixed quantities rather than random variables, meaning that ascribing probabilities to them being contained within an interval does not make sense.

In contrast to the above, a Bayesian interpretation views probability as a *degree of belief*; given the available information, how much do I believe a proposition to be true? Unlike in a frequentist interpretation, parameters are random variables that can be assigned probability distributions which quantify our uncertainty about their value.

This notion of the available information drives many of the practical differences between Bayesian and frequentist inference. This information includes the data that was observed, but makes no reference to any infinite repetitions of experiments not performed. This property has been seen as a significant philosophical advantage of Bayesian methods (e.g., Trotta, 2008). The available information also includes prior beliefs over the value of the parameters before any data has been observed, e.g., from the results of previous experiments or physical restrictions on parameters such as masses being positive (e.g., Gelman, 2006).

It is also possible to construct intervals from the results of Bayesian inference, known as credible intervals. Unlike their frequentist analogues, these intervals can be

interpreted to be intervals that contain the true parameter with a specified probability. This property is an example of what is seen (e.g., Jaynes and Kempthorne, 1976; Jaynes, 2003; Briggs, 2012) as an advantage of Bayesian methods; their interpretation in the context of data analysis is clear. Frequentist methods, through their use of methods such as maximum-likelihood estimation, provide results that quantify the probability of the data. The usefulness of this has been questioned; Trotta (2008), for example, argues that a statistical analysis should depend only on the data that have been observed, not on hypothetical datasets that *could* have been. The interpretation of frequentist probabilities also necessitates having a repeatable experiment (Sprott, 2008), a potentially problematic quality for the analysis of cosmological data where only one sky can be observed.

Alternatively, Bayesian methods allow the probability of parameters and models, conditional on the observed data and an explicit set of assumptions, to be obtained directly. These quantities are those that are of interest for scientific enquiry. A Bayesian interpretation of probability also makes sense irrespective of the number of samples observed. Bayesian inference therefore provides a natural framework for statistical analysis of cosmological data. The remainder of section 3.1 describes the theory and application of Bayesian inference in detail.

3.1.2 Probability theory

Mathematical investigations of probability have a long history with notable contributions by Pascal and Fermat in the 17^{th} century (Ore, 1960), Bernoulli (1713) and Laplace (1820). However, our modern understanding of probability theory was first introduced by Kolmogorov (1933), which put probability in the context of measure theory.

An alternative formulation of probability theory is presented by Cox (1946) and discussed in detail by Jaynes (2003). Cox considered probability in terms of logical reasoning, extending the binary true-or-false logic of Boole (1847) to encompass uncertainty by moving to continuous values $\in [0, 1]$. The Cox-Jaynes theorem derives the results of probability theory from three *desiderata*. Importantly for a Bayesian interpretation of probability, this is done without appealing to the notion of repeated trials, as would be needed for a frequentist formulation of probability. The theorem is therefore a justification that the rules of probability theory are compatible with a Bayesian interpretation of probability as degrees of belief.

The first *desideratum* of the Cox-Jaynes theorem is that the plausibility of an proposition should be represented by a real number. Secondly, this plausibility should

behave in a way that follows common sense. This is presented by Jaynes (2003) in terms of conditional reasoning; that is, the plausibility of one proposition, given that another proposition is known to be true.

Consider three propositions, A, B and C. Also assume that we know the plausibility of both A given C, and of B given A and C. Borrowing the standard notation of conditional probabilities, we write these as $P(A \mid C)$ and $P(B \mid A, C)$ respectively. We then consider the effect of updating proposition $C \to C'$ such that proposition A is more plausible, i.e.,

$$P(A \mid C') > P(A \mid C)\,, \tag{3.3}$$

but that the plausibility of proposition B remains the same, i.e.,

$$P(B \mid A, C') = P(B \mid A, C)\,. \tag{3.4}$$

We now consider the effect on the plausibility of both A and B together, given C, after this update. Since A has become more plausible and B has not changed, the common sense expectation is that this should not become less plausible, i.e., we should expect that

$$P(A, B \mid C') \geq P(A, B \mid C)\,. \tag{3.5}$$

The final *desideratum* is that plausibilities should behave consistently. This is defined by Jaynes (2003) to mean that equal plausibilities should be assigned to equal states of knowledge, these plausibilities should be based on all available knowledge, and the same conclusions reasoned in several different ways should be equally plausible.

From these *desiderata*, the Cox-Jaynes theorem derives the following two rules of probability theory, the sum rule

$$P(A \mid B) + P(\bar{A} \mid B) = 1\,, \tag{3.6}$$

where $\bar{A}$ denotes the proposition that A is false, and the product rule

$$P(A, B \mid C) = P(A \mid B, C)P(B \mid C)\,. \tag{3.7}$$

The derivation here considered $P(A)$ to be the probability of a logical proposition A. However, the probability distributions we will consider throughout this thesis will be functions of real values and will thus be of the form $P(X = x)$, i.e., the probability of the logical proposition that the random variable X has the value x. Throughout, we will notate this simply as $P(x)$. When x can take only discrete values, these functions are known as *probability mass functions* (PMFs). When x is continuous, the functions

are known as *probability density functions* (PDFs).

Various results for the manipulation of probabilities can be derived from the sum and product rules above. These are detailed in the following sections.

3.1.3 Bayes' theorem

The central result on which the rest of Bayesian inference is based is Bayes' theorem, a rule regarding conditional probabilities named after and first considered by Bayes (1763), and subsequently rederived independently by Laplace (1829). The rule follows simply from the product rule defined in equation 3.7 by noting that the joint probability $P(A, B)$ is trivially equal to that of $P(B, A)$. Replacing the propositions A, B and C with $\boldsymbol{\theta}$, $\boldsymbol{d}$ and $\mathcal{M}$ respectively, this gives

$$P(\boldsymbol{\theta}, \boldsymbol{d} \mid \mathcal{M}) = P(\boldsymbol{\theta} \mid \boldsymbol{d}, \mathcal{M}) P(\boldsymbol{d} \mid \mathcal{M}) = P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M}) P(\boldsymbol{\theta} \mid \mathcal{M}) \,. \tag{3.8}$$

A simple rearranging of this equation gives the expression for Bayes' theorem,

$$P(\boldsymbol{\theta} \mid \boldsymbol{d}, \mathcal{M}) = \frac{P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M}) P(\boldsymbol{\theta} \mid \mathcal{M})}{P(\boldsymbol{d} \mid \mathcal{M})} \,. \tag{3.9}$$

Interpreting $\boldsymbol{\theta}$ as the parameters of interest for a model $\mathcal{M}$ under consideration, and $\boldsymbol{d}$ as the observed data from which we wish to make inferences, we can identify the four probabilities present in this equation.

Firstly, $P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M})$ is the likelihood, the probability of the data, given the value of the parameters $\boldsymbol{\theta}$ under the model $\mathcal{M}$. This is the same definition as the likelihood we discussed previously in the context of frequentist statistics, though we have made the conditioning more explicit here.

Secondly, $P(\boldsymbol{\theta} \mid \mathcal{M})$ is the prior distribution. As discussed in section 3.1.1, this is our belief for the distribution of the parameter before seeing any data, and is therefore an explicit representation of our modelling assumptions. This term can be informed by previous experiments or physical arguments, but could also represent a state of ignorance; this is discussed in more detail in section 3.1.7.

Thirdly, $P(\boldsymbol{d} \mid \mathcal{M})$ is the evidence, or the marginal likelihood. This is the normalising constant for Bayes' theorem, ensuring that the left-hand side of equation 3.9 describes a valid probability density that integrates w.r.t. $\boldsymbol{\theta}$ to unity. The role of this term in parameter inference and model comparison is discussed in sections 3.1.5 and 3.1.6 respectively.

Lastly, $P(\boldsymbol{\theta} \mid \boldsymbol{d}, \mathcal{M})$ is the posterior distribution. This is the distribution defined by Bayes' theorem and quantifies our beliefs about the parameters $\boldsymbol{\theta}$ of model $\mathcal{M}$ after incorporating the information gained from the observed data $\boldsymbol{d}$. Constructing this distribution is typically the main goal of modelling in a Bayesian manner; inference tasks such as parameter estimation and model comparison are completed through the manipulation or utilisation of this posterior distribution in some way. The details of how these tasks are performed are detailed in the following sections.

Inspecting equation 3.9, we see that the effect of Bayes' theorem is to reverse the order of conditioning. The importance of this can be seen by considering the definitions above. The likelihood describes the forward model, the generative process by which data can be obtained from the specified model; different values of the parameters $\boldsymbol{\theta}$ will produce different distributions of data. However, as detailed in section 3.1.1, the goal of statistical inference is to use observed data to obtain information about the data-generating process, described by the model and the parameters that control it. This is known as the *inverse problem*. The posterior distribution is the result of this inverse problem. Bayesian inference thus mathematically describes the logical procedure of modifying existing beliefs, encoded in the prior, to take account of new data. This process in known as *Bayesian updating*.

3.1.4 Marginalisation

Another useful property that can be derived from the results of section 3.1.2 is marginalisation. We follow here the derivation detailed in Sivia and Skilling (2006). First, consider the sum of the two joint distributions $P(\theta, \phi \mid d)$ and $P(\theta, \bar{\phi} \mid d)$, where ϕ denotes a proposition that can take the values of true or false, and the bar denotes the negation of this value, as before. Expanding the terms in this sum using the product rule of equation 3.7, this becomes

$$\begin{aligned} P(\theta, \phi \mid d) + P(\theta, \bar{\phi} \mid d) &= P(\phi \mid \theta, d)P(\theta \mid d) + P(\bar{\phi} \mid \theta, d)P(\theta \mid d) \\ &= \left[P(\phi \mid \theta, d) + P(\bar{\phi} \mid \theta, d)\right] P(\theta \mid d)\,. \end{aligned} \tag{3.10}$$

From the sum rule defined in equation 3.6, the terms in the square brackets sum to unity. Thus,

$$P(\theta \mid d) = P(\theta, \phi \mid d) + P(\theta, \bar{\phi} \mid d)\,. \tag{3.11}$$

It is possible to generalise this statement by considering a proposition that can take one of several values $\{\phi_i\} = \{\phi_1, \phi_2 \ldots \phi_N\}$ rather than simply true or false. Assume that the set $\{\phi_i\}$ is both *exclusive*, meaning that two elements cannot be true

at the same time, and *exhaustive*, meaning that the proposition can take no other values; thus, the set can have exactly one true element. We can then write a sum over joint distributions analogously to equation 3.10, giving

$$\begin{aligned}\sum_{i=1}^{N} P(\theta, \phi_i \mid d) &= P(\theta, \phi_1 \mid d) + P(\theta, \phi_2 \mid d) \cdots + P(\theta, \phi_N \mid d) \\ &= \left[P(\phi_1 \mid \theta, d) + P(\phi_2 \mid \theta, d) \cdots + P(\phi_N \mid \theta, d)\right] P(\theta \mid d) \\ &= \left[\sum_{i=1}^{N} P(\phi_i \mid \theta, d)\right] P(\theta \mid d)\,.\end{aligned} \tag{3.12}$$

Following from the assumptions above, the sum in the square brackets is equal to unity. Thus, we find that

$$P(\theta \mid d) = \sum_{i=1}^{N} P(\theta, \phi_i \mid d)\,. \tag{3.13}$$

This is simply the law of total probability (e.g., Kokoska and Zwillinger, 2000). In the continuum limit, this expression can be written as an integral,

$$P(\theta \mid d) = \int P(\theta, \phi \mid d)\, \mathrm{d}\phi\,, \tag{3.14}$$

giving a property that applies to a continuous parameter ϕ. This process allows an additional parameter to be incorporated into a probability distribution and is known as marginalisation. This can be useful when the forward model for a system can only be written in terms of a parameter that is unknown but is not of interest, referred to as a *nuisance parameter*. Marginalisation accounts for the uncertainty in nuisance parameters, propagating it correctly into the posterior of the parameters of interest. If these unknown parameters were instead fixed to an *a priori* reasonable value, the resulting posterior would be overly narrow and not an accurate representation of the true state of knowledge about the system.

3.1.5 Parameter inference

The most common application of Bayes rule is to the problem of parameter inference — for a particular fixed model, what constraints can we put on its parameters, given a dataset? A typical cosmological example of this procedure is the placing of constraints on the parameters of the ΛCDM model. The dataset may be comprised of a single cosmological probe such as weak lensing (e.g., Köhlinger et al., 2017), or of a collection of multiple probes (e.g., Krause et al., 2017). This change in the dataset would manifest in a different forward model defined in likelihood. The joint likelihood for multiple

independent probes can be constructed simply as the product of the individual likelihood. However, correlated datasets would require a likelihood that accounts for this correlation.

When in the context of performing parameter inference, Bayes' theorem is often written in a simplified form, given by

$$P(\boldsymbol{\theta} \mid \boldsymbol{d}) \propto P(\boldsymbol{d} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) \,. \tag{3.15}$$

There are two simplifications here from the full expression given in equation 3.9. Firstly, the model conditioning on the model has been removed. The posterior distribution is still implicitly conditioned on the choice of model, but this choice is fixed when performing parameter inference. Therefore, the model conditioning is typically suppressed for clarity.

Secondly, the evidence term in the denominator has been removed. As a result, the expression on the right hand side no longer integrates to unity and so does not describe a valid probability density. However, this is sufficient for the purpose of parameter inference. In this case, both the model and the dataset are fixed, with only the value of the parameters varying. Since the evidence $P(\boldsymbol{d} \mid \mathcal{M})$ depends only on the data and the model, it is constant and can be neglected.

Parameter inference can now be performed. This is typically done by drawing samples from the posterior. With the removal of the evidence, the posterior distribution is only proportional to this expression. In this case, parameter sampling must therefore be performed by a method that is insensitive to the overall normalisation of the probability density, such as Markov chain Monte Carlo (MCMC). Sampling methods are discussed in more detail in in section 3.1.9.

For a restricted set of combinations of likelihood and prior choices known as *conjugate distributions*, the posterior distribution can be found analytically, avoiding the need for sampling. This is discussed in more detail in section 3.1.7.

3.1.6 Model comparison

In addition to making inferences about parameters of a fixed model, Bayesian inference also provides a method of comparing the relative probability of two models given a dataset. Bayesian model comparison can be performed by applying Bayes' theorem to obtain the posterior distribution of a model itself, rather than its parameters. The

posterior for model $\mathcal{M}_i$ is then given by

$$P(\mathcal{M}_i \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \mathcal{M}_i)P(\mathcal{M}_i)}{P(\boldsymbol{d})}\,. \tag{3.16}$$

It is helpful to rewrite the denominator of this expression using the marginalisation property introduced in section 3.1.4, giving

$$P(\mathcal{M}_i \mid \boldsymbol{d}) = \frac{P(\boldsymbol{d} \mid \mathcal{M}_i)P(\mathcal{M}_i)}{\sum_i P(\boldsymbol{d} \mid \mathcal{M}_i)P(\mathcal{M}_i)}\,. \tag{3.17}$$

In principle, this sum runs over all possible choice of model. In practice, however, it is not generally possible to define the set of all possible models to marginalise over[1]. As a result, the absolute probability of a model cannot be evaluated in this way.

Bayesian model comparison instead proceeds by comparing the relative probability of two competing models. This ratio between models $\mathcal{M}_1$ and $\mathcal{M}_2$ is given by

$$\frac{P(\mathcal{M}_1 \mid \boldsymbol{d})}{P(\mathcal{M}_2 \mid \boldsymbol{d})} = \frac{P(\boldsymbol{d} \mid \mathcal{M}_1)}{P(\boldsymbol{d} \mid \mathcal{M}_2)}\frac{P(\mathcal{M}_1)}{P(\mathcal{M}_2)}\,. \tag{3.18}$$

Since the problematic denominator has cancelled, this quantity can now be successfully evaluated. The above expression contains two fractions, the ratio of the model priors, and the ratio of the evidences under each model, typically referred to as the Bayes factor

$$\mathcal{B}_{1,2} \equiv \frac{P(\boldsymbol{d} \mid \mathcal{M}_1)}{P(\boldsymbol{d} \mid \mathcal{M}_2)}\,. \tag{3.19}$$

When the ratio of model priors is unity, i.e., neither model is preferred *a priori*, the Bayes factor quantifies the degree to which the data supports model $\mathcal{M}_1$ over $\mathcal{M}_2$. A commonly cited qualitative interpretation of this quantity was provided by Jeffreys (1939), eponymously referred to as the *Jeffreys' scale*.

The Jeffreys' scale interpretation is given in Table 3.1. When the log-Bayes factor is positive as in the table, this indicates a preference for model $\mathcal{M}_1$. Negative log-Bayes factors are qualitatively interpreted in the same way, but instead indicate a preference for model $\mathcal{M}_2$.

As detailed in section 3.1.5, the evidence term $P(\boldsymbol{d} \mid \mathcal{M})$ is typically neglected during parameter inference. However, due to its existence in the Bayes factor, the evidence is an integral part of Bayesian model selection. In order to see the practical requirements of computing this term, it is helpful to rewrite it including the parameter

[1]This sum can be restricted to a specific, finite set of models by assumption if it is required that the posterior is normalised, e.g., for the purpose of model averaging. This is discussed later in this section.

Table 3.1: A summary of the Jeffreys' scale, a qualitative interpretation of Bayes factors, directly quoted from Jeffreys (1939). The positive log-values here indicate a preference in favour of model $\mathcal{M}_1$. For negative log-values, the qualitative descriptions are the same, but are instead in favour of model $\mathcal{M}_2$.

Bayes factor	Description
$0 < \log_{10}(\mathcal{B}_{1,2}) < 1/2$	Not worth more than a bare mention
$1/2 < \log_{10}(\mathcal{B}_{1,2}) < 1$	Substantial
$1 < \log_{10}(\mathcal{B}_{1,2}) < 3/2$	Strong
$3/2 < \log_{10}(\mathcal{B}_{1,2}) < 2$	Very strong
$\log_{10}(\mathcal{B}_{1,2}) > 2$	Decisive

vector $\boldsymbol{\theta}$ by marginalising, and factorising using product rule. Doing this, we find

$$P(\boldsymbol{d} \mid \mathcal{M}) = \int P(\boldsymbol{d}, \boldsymbol{\theta} \mid \mathcal{M}) \, \mathrm{d}\boldsymbol{\theta} = \int P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M}) P(\boldsymbol{\theta} \mid \mathcal{M}) \, \mathrm{d}\boldsymbol{\theta} \,. \tag{3.20}$$

The two resulting terms are the likelihood and the prior. Thus, this integrand is the numerator of Bayes' theorem, the unnormalised density defined in equation 3.15 that is typically used for Bayesian parameter inference problems. Expanding the evidence in this way shows why it correctly normalised Bayes' theorem, but also reveals what is required to evaluate it; the unnormalised posterior must be integrated over the entirety of its support in parameter space.

In practice, this integral can often be difficult to evaluate, particularly if the dimensionality of the parameter space is large. While the prior volume may be large, the likelihood can peak sharply. Nevertheless, the comparatively low-density tails of the posterior can contain significant volume and can therefore not be ignored. Numerically evaluating an integral with non-negligible contributions at both of these scales is computationally challenging. Instead, methods such as nested sampling (Skilling, 2006) that evaluate this integral efficiently have been developed; this is discussed in more detail in section 3.1.9.

One convenient exception to this computational expense is the case of *nested models*. A pair of nested models consists of a simple model with likelihood $P_1(\boldsymbol{d} \mid \boldsymbol{\theta})$ and prior $P_1(\boldsymbol{\theta})$, and a complex model with likelihood $P_2(\boldsymbol{d} \mid \boldsymbol{\theta}, \boldsymbol{\phi})$ and prior $P_2(\boldsymbol{\theta}, \boldsymbol{\phi})$. The models are considered nested if the complex model contains the simple model; that is, at a specific value of the parameter $\boldsymbol{\phi} = \boldsymbol{\phi}_0$, its likelihood $P_2(\boldsymbol{d} \mid \boldsymbol{\theta}, \boldsymbol{\phi}_0) = P_1(\boldsymbol{d} \mid \boldsymbol{\theta})$ equals that of the simpler model, and its prior $P_2(\boldsymbol{\theta} \mid \boldsymbol{\phi}_0) = P_1(\boldsymbol{\theta})$. In this case, the

Savage-Dickey density ratio (Dickey, 1971) allows the Bayes factor to be evaluated as

$$\mathcal{B}_{1,2} = \frac{P_2(\boldsymbol{\phi}_0 \mid \boldsymbol{d})}{P_2(\boldsymbol{\phi}_0)}\,, \tag{3.21}$$

where the numerator is the marginalised posterior of the complex model evaluated at $\boldsymbol{\phi} = \boldsymbol{\phi}_0$, and the denominator is the marginalised prior evaluated at the same point.

Any method of model comparison must balance a models ability to predict the observed data with its complexity. A model can be made arbitrarily complex by giving it an increasingly large number of parameters. As a result, this model would be extremely flexible, in the sense that it would be able to predict many different datasets. This is summarised in a quote by John von Neumann (e.g., Dyson, 2004), that *"with four parameters I can fit an elephant, and with five I can make him wiggle his trunk"* [2]. We should therefore not be surprised that a model with many parameters explains the data better than one with few; this increase in the goodness-of-fit should be balanced against the corresponding increase in model complexity.

This notion of balancing a models complexity with its goodness-of-fit is known as *Occam's razor*; given two models that make the same prediction, the simpler model should be preferred. An advantage of Bayesian model comparison is that it automatically includes this effect (e.g., Jefferys and Berger, 1992). As the complexity of a model $\mathcal{M}$ grows, it is able to predict a wider variety of datasets. As a result, the evidence $P(\boldsymbol{d} \mid \mathcal{M})$ evaluated given an observed dataset $\boldsymbol{d}$ will be reduced. The goodness-of-fit, encoded in the likelihood $P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M})$, must therefore increase to compensate this if the more complex model is to be preferred.

It is important to note that, for the purpose of model comparison, the prior distribution $P(\boldsymbol{\theta} \mid \mathcal{M})$ must be correctly normalised, i.e., $\int P(\boldsymbol{\theta} \mid \mathcal{M})\ \mathrm{d}\boldsymbol{\theta} = 1$. As a result, this excludes the use of improper priors; see section 3.1.7 for details. In addition, the evidence is sensitive to the choice of prior, even when the likelihood is strongly peaked.

Model averaging

An additional use of the evidence term is model averaging (e.g., Fragoso et al., 2018), where the choice of model is marginalised over. Thus, the posterior for a parameter

[2] Indeed, if one permits complex parameters, this has been shown to be true (Mayer et al., 2010).

vector $\boldsymbol{\theta}$ averaged over a possible N models is given by

$$\begin{aligned} P(\boldsymbol{\theta} \mid \boldsymbol{d}) &= \sum_{i=1}^{N} P(\boldsymbol{\theta}, \mathcal{M}_i \mid \boldsymbol{d}) \\ &= \sum_{i=1}^{N} P(\boldsymbol{\theta} \mid \boldsymbol{d}, \mathcal{M}_i) P(\mathcal{M}_i \mid \boldsymbol{d}) \,, \end{aligned} \tag{3.22}$$

where $P(\boldsymbol{\theta} \mid \boldsymbol{d}, \mathcal{M}_i)$ is the posterior under model $\mathcal{M}_i$, and $P(\mathcal{M}_i \mid \boldsymbol{d})$ is the *a posteriori* probability for that model, as defined in equation 3.17.

Implicitly, this assumes that the set of models $\{\mathcal{M}_i\} = \{\mathcal{M}_1, \mathcal{M}_2 \ldots \mathcal{M}_N\}$ is exhaustive, meaning that the evidences are normalised $\sum_{i=1}^{N} P(\mathcal{M}_i \mid \boldsymbol{d}) = 1$. Marginalising over the choice of model in this way incorporates its uncertainty into the final posterior over the parameters of interest.

3.1.7 Priors

An integral part of a Bayesian analysis, and one that significantly differentiates it from a frequentist analysis, is the choice of $P(\boldsymbol{\theta} \mid \mathcal{M})$, the prior distribution. This distribution quantifies the beliefs about the parameters $\boldsymbol{\theta}$ before the data has been seen. This belief may take the form of the results from a previous experiment determining the same quantity, leading to the adage that *"yesterday's posterior is today's prior"* (Lindley, 2000), or from additional data that constrains properties of a population the system of interest is known to have been drawn from. Alternatively, theoretical considerations can inform a prior distribution, e.g., a nonnegative prior that constrains a mass $m > 0$. Such priors are typically referred to as *subjective priors*.

In contrast to these, priors can be constructed for a particular problem which aim to minimise their information content. Priors chosen through such a mathematical procedure are termed *objective priors*. Examples of these are discussed below.

Proper and improper priors

Before we introduce different types of priors, it is worth briefly discussing the difference between proper and improper priors. For a probability distribution $P(\boldsymbol{x})$ to be valid, it must be normalised, i.e., $\int P(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = 1$. Priors that are valid in this way are known as *proper* priors.

If the integral of a prior distribution is not finite, it cannot be normalised and is referred to as *improper* (e.g., Sivia and Skilling, 2006). Since model comparison

requires proper priors that are correctly normalised, improper priors cannot be used for this purpose. In addition, the posterior is not guaranteed to be proper when the prior is not. Since an improper posterior is not a valid probability distribution, the use of an improper prior necessitates checking the propriety of the posterior.

Objective priors

If nothing is known about a parameter, a seemingly reasonable choice of an uninformative prior might be a uniform prior, i.e., $P(\boldsymbol{\theta} \mid \mathcal{M}) \propto 1$. In fact, the first application of a Bayesian analysis by Bayes (1763) involved inference of the probability of success p from a series of Bernoulli success-or-failure trials with a uniform prior over p. However, there exist several problems with this choice.

Firstly, a uniform prior without lower or upper bounds does not have a finite integral. Inference must therefore proceed with the caveats above. A uniform prior can be made proper by imposing lower and upper bounds outside which the prior is zero. Note, however, that this prior is no longer uninformative, since this imposes bounds on the possible values of the parameters. Secondly, a uniform prior can be parametrisation-dependent. This means that the change-of-variables from $\boldsymbol{\theta}$ where $P(\boldsymbol{\theta})$ is uninformative uniform prior to $\boldsymbol{\phi}$ can induce a prior $P(\boldsymbol{\phi})$ which is highly informative.

Instead of a uniform prior, methods exist to mathematically construct priors that are uninformative. An early example of this is presented by Jeffreys (1939), where the prior on a parameter θ is given by

$$P(\theta) \propto \sqrt{I(\theta)}\,, \tag{3.23}$$

where $I(\theta)$ is the Fisher information (e.g., Ly et al., 2017), given by

$$I(\theta) = -\int \left(\frac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log P(\boldsymbol{d} \mid \theta, \mathcal{M}) \right) P(\boldsymbol{d} \mid \theta, \mathcal{M}) \, \mathrm{d}\boldsymbol{d}\,. \tag{3.24}$$

An alternative approach to objective priors are *reference priors* (Bernardo, 1979), a information theory-based method that precisely defines the notion of a uninformative prior. In particular, reference priors are constructed by maximising the gain in information during the Bayesian update from prior to posterior. The information gain is measured through the Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951), given by

$$D_{\mathrm{KL}}(P \mid Q) = \int P(\boldsymbol{\theta}) \log \frac{P(\boldsymbol{\theta})}{Q(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta}\,. \tag{3.25}$$

Since the prior distribution should quantify the belief about the parameters before seeing any data, the true posterior, which depends on the observed data, cannot be used to construct the reference prior. Instead, the maximised quantity is the *expected information*, the expectation value of the KL divergence between the prior and posterior, averaged over the data distribution, i.e.,

$$\mathcal{I} = \int P(\boldsymbol{d}) \left[\int P(\boldsymbol{\theta} \mid \boldsymbol{d}) \log \frac{P(\boldsymbol{\theta} \mid \boldsymbol{d})}{P(\boldsymbol{\theta})} \, \mathrm{d}\boldsymbol{\theta} \right] \, \mathrm{d}\boldsymbol{d} \,, \tag{3.26}$$

where $P(\boldsymbol{d}) = \int P(\boldsymbol{d} \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}$. The reference prior is then the distribution $P(\boldsymbol{\theta})$ that maximises equation 3.26. When this procedure is applied to one-dimensional models, the reference prior is equal to the Jeffrey's prior, though this is not the case in general for multivariate models (Berger et al., 2009). This objective Bayesian approach has found use in several astrophysical inference problems (e.g., Knoetig, 2014; Heavens and Sellentin, 2018; Jew and Grumitt, 2019).

Conjugate priors

A particularly mathematically convenient type of prior is known as *conjugate priors*. When paired with the appropriate likelihood to make a pair of *conjugate distributions*, the posteriors of these models have the property that they are the same probability distribution as the prior but with different parameters. These posterior distributions can therefore be written in closed form. An advantage of this is that some *a posteriori* properties of the parameters, such as their mean and variance, can also be found analytically, provided that closed form expressions exist for these properties for the relevant probability distribution. Constructing models in this way can therefore avoid the need for sampling, at the expense of restricting the variety of models possible to those involving conjugate distributions.

3.1.8 Bayesian hierarchical modelling

An advantage of Bayesian inference is the ease with which uncertainties can be propagated rigorously, including correlations between parameters. These uncertainties are encompassed within a probability distribution. These distributions can be the final result of inference for the parameters of interest, but can also represent intermediate products not directly of interest, but required for the final inference. Complex models can be built in this way, assigning probability distributions to all unknown intermediate quantities and conditioning on these quantities elsewhere. These are known as *Bayesian hierarchical models* (BHMs).

As a toy example, consider a system where we observe a dataset of N samples $\{\boldsymbol{d}\} = \{\boldsymbol{d}_1, \boldsymbol{d}_2 \dots \boldsymbol{d}_N\}$ independently drawn from a population. Each sample i has associated parameters $\boldsymbol{\theta}_i$ that govern the data generating process described by the individual likelihood $P(\boldsymbol{d}_i \mid \boldsymbol{\theta}_i)$. From the assumption of independence, the total likelihood is given by the product over individual likelihoods, i.e.,

$$P(\{\boldsymbol{d}\} \mid \{\boldsymbol{\theta}\}) = \prod_{i=1}^{N} P(\boldsymbol{d}_i \mid \boldsymbol{\theta}_i)\,, \tag{3.27}$$

where $\{\boldsymbol{\theta}\} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \dots \boldsymbol{\theta}_N\}$ is the set of individual sample parameters. In a simple Bayesian analysis, the posterior distribution would then simply be given by

$$P(\{\boldsymbol{\theta}\} \mid \{\boldsymbol{d}\}) \propto \prod_{i=1}^{N} P(\boldsymbol{d}_i \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i)\,, \tag{3.28}$$

where the total prior $P(\{\boldsymbol{\theta}\})$ has been separated by the assumption of independence.

This approach requires that a prior can be written down. If the population from which each sample has been drawn is well understood, this can be done. However, for many complex problems, this is a strong requirement. Instead, a desirable property of the analysis is that we could make inferences not only about the individual samples drawn from the population, but also about the population itself. To do this, we need to incorporate the prior information that these samples are all drawn from the same population. This can be through the structure of the prior by constructing a hierarchical model.

To do this, we parametrise the prior distribution with its own set of parameters $\boldsymbol{\phi}$, known as *hyperparameters*, giving $P(\boldsymbol{\theta}_i \mid \boldsymbol{\phi})$. Unlike the sample parameters $\boldsymbol{\theta}$, these hyperparameters are shared across the entire population. The hyperparameters are also unknown quantities, meaning we must also place priors over these hyperparameters $P(\boldsymbol{\phi})$, referred to as *hyperpriors*. These hyperparameters are then included as part of the final posterior distribution. Applying Bayes rule, this is given by

$$P(\{\boldsymbol{\theta}\}, \boldsymbol{\phi} \mid \{\boldsymbol{d}\}) \propto P(\{\boldsymbol{d}\} \mid \{\boldsymbol{\theta}\}, \boldsymbol{\phi}) P(\{\boldsymbol{\theta}\}, \boldsymbol{\phi})\,. \tag{3.29}$$

A common assumption when constructing BHMs is that terms are dependent only on the quantities directly above them in the hierarchy, i.e., that the likelihood term is given by equation 3.27, as before. Inserting this and separating the joint prior using

product rule, the posterior becomes

$$P(\{\boldsymbol{\theta}\}, \boldsymbol{\phi} \mid \{\boldsymbol{d}\}) \propto P(\boldsymbol{\phi}) \prod_{i=1}^{N} P(\boldsymbol{d}_i \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}) \,. \tag{3.30}$$

Note that, since the hyperparameters are shared across all samples, the hyperprior has been moved outside of the product.

This hierarchical model now allows observations of samples from the population to be used to make inferences about the population itself by placing a posterior distribution over its parameters $\boldsymbol{\phi}$. If these population parameters are not of interest, the posterior distribution over the sample parameters only can be obtained simply by marginalising, i.e.,

$$P(\{\boldsymbol{\theta}\} \mid \{\boldsymbol{d}\}) \propto \int P(\boldsymbol{\phi}) \prod_{i=1}^{N} P(\boldsymbol{d}_i \mid \boldsymbol{\theta}_i) P(\boldsymbol{\theta}_i \mid \boldsymbol{\phi}) \, \mathrm{d}\boldsymbol{\phi} \,. \tag{3.31}$$

Doing this incorporates the uncertainty on the intermediate $\boldsymbol{\phi}$ parameters into the final posterior distribution. In addition to providing a rigorous way to propagate uncertainties, the hierarchical structure of BHMs also provides a way for inferences about individual samples to be informed by one another, meaning that that width of the marginal $\boldsymbol{\theta}_i$ posteriors will be reduced, tightening the constraints. This effect is known as *shrinkage* (e.g., Gelman et al., 2013). The assumption of independence and lack of any global structure means that this does not occur in the non-hierarchical model above.

BHMs are often presented as *directed acyclic graphs* (DAGs), diagrams which show the conditional dependence between terms in the model. As described in section 3.1.9, this representation can be useful for obtaining the conditional distributions required for Gibbs sampling. The DAGs of the toy model and its hierarchical alternative above are shown in Figure 3.1.

Bayesian hierarchical modelling has been used for several cosmological data analysis problems. For example, Alsing et al. (2016) construct a BHM over both the cosmic shear power spectrum and the pixelised cosmic shear map, avoiding difficulties created by masked survey areas. This method was applied to Canada-France-Hawaii Telescope Lensing Survey (CFHTLens, Heymans et al., 2012) data in Alsing et al. (2017). Leistedt and Hogg (2017) derive a BHM of the colour-magnitude diagram from Gaia (Gaia Collaboration et al., 2016) parallax and magnitude measurements in order to infer stellar distances. They find a significant improvement in the precision of distance inferences over non-hierarchical approaches. Feeney et al. (2018) use a BHM describing

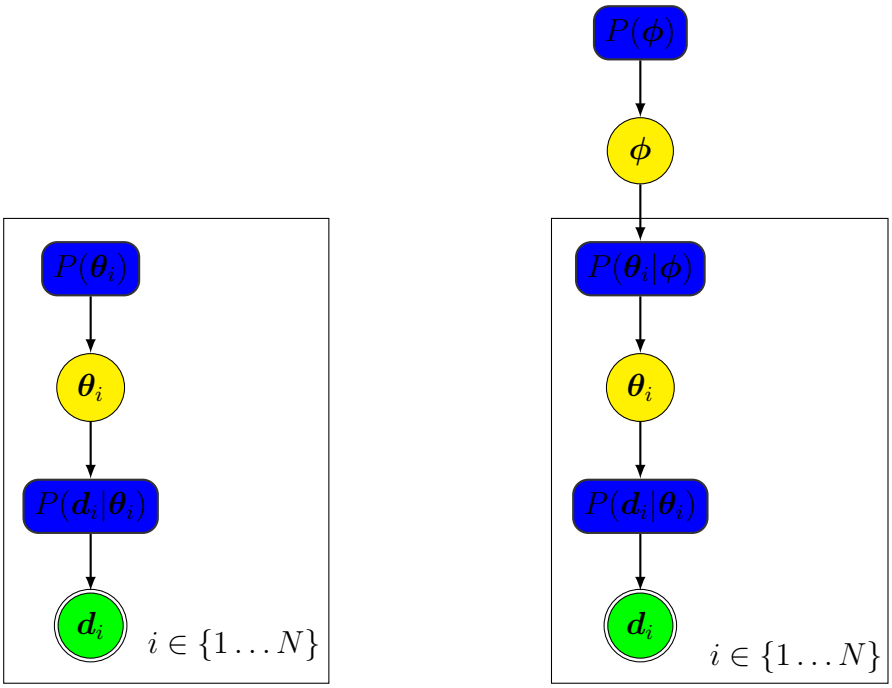

Figure 3.1: Directed acyclic graphs for the two models described in section 3.1.8. The graph on the left shows the simple model defined in equation 3.28, while the graph on the right shows the hierarchical alternative defined in equation 3.30. Terms in the blue rounded rectangles are probability distributions. Quantities in single circles are latent parameters sampled from probability distributions, and those in double circles are observed quantities. The black rectangles indicate a product over independent terms.

the local distance ladder in order to rigorously propagate uncertainties from all observations and quantify the tension between local and cosmological measurements of the Hubble parameter; see section 2.3.4 for a discussion of the Hubble parameter tension. Using Bayesian model comparison, they find that the odds against ΛCDM are reduced compared to a naive comparison of the measurements, but that a tension still remains. Finally, Leistedt et al. (2016) use a BHM to jointly infer the redshifts of each galaxy in a population and the redshift distribution of the population itself using photometric observations. This method is discussed in more detail in section 4.2.1

3.1.9 Sampling probability distributions

As discussed in section 3.1.5, making inferences from a posterior that has been derived using Bayes' theorem typically involves sampling from that distribution. In general, quantities of interest cannot be obtained in closed form from the posterior distribution; the notable exception to this are models formed from conjugate distributions as described in section 3.1.7. In addition, the evidence $P(\boldsymbol{d} \mid \mathcal{M})$ involves an integration over the entire support of the posterior in parameter space. As discussed in section 3.1.6, this integral is often computationally difficult.

Instead, a set of samples drawn from the posterior distribution are extremely useful for inference. The simplest thing to do with samples is to plot them. By plotting histograms of the samples, posterior distributions can be inspected, an important sanity check. Sampling a posterior distribution also allows variables to be easily marginalised out. Given a set of samples $\boldsymbol{\theta}_i, \boldsymbol{\phi}_i \sim P(\boldsymbol{\theta}, \boldsymbol{\phi} \mid \boldsymbol{d})$ for $i \in \{1 \dots N\}$, discarding the unwanted part of the samples $\{\phi\}$ leaves a set of samples $\{\boldsymbol{\theta}\}$ drawn from the marginal posterior $P(\boldsymbol{\theta} \mid \boldsymbol{d})$.

It can sometimes be useful to summarise the results of a posterior distribution as a single number, known as a *point estimate*[3]. A common way to do this is to calculate expectation values, defined as

$$\mathrm{E}[\boldsymbol{\theta}] \equiv \int \boldsymbol{\theta} P(\boldsymbol{\theta} \mid \boldsymbol{d}) \, \mathrm{d}\boldsymbol{\theta} \,. \tag{3.32}$$

This expectation is sometimes labelled $\mathrm{E}[\boldsymbol{\theta} \mid \boldsymbol{d}]$ to explicitly denote the conditioning on the data $\boldsymbol{d}$. This can be generalised to the expectation value of a deterministic

[3]Note that this compression will inevitably result in a loss of information. A point estimate alone has no uncertainty associated with it, while a point estimate with error bars must make assumptions about the shape of the posterior distribution.

function of a random variable $f(\boldsymbol{\theta})$ by

$$\mathrm{E}[f(\boldsymbol{\theta})] \equiv \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta} \mid \boldsymbol{d}) \, \mathrm{d}\boldsymbol{\theta} \,. \tag{3.33}$$

Given a set of N samples drawn from the posterior $\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta} \mid \boldsymbol{d})$ for $i \in \{1 \ldots N\}$, the expectation defined in equation 3.33 can be approximated as

$$\mathrm{E}[f(\boldsymbol{\theta})] \approx \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta}_i) \,. \tag{3.34}$$

Thus, the mean of samples drawn from the posterior is the expectation value.

An important property for many sampling methods is that they do not require the density from which they sample to be normalised. This obviates the need to calculate the evidence for parameter estimation, potentially reducing the computational burden significantly. For model comparison problems where calculating the evidence is unavoidable, sampling methods exist that can efficiently evaluate this integral. A variety of these sampling methods are discussed throughout the rest of this section.

Pseudorandom number generation

Before we discuss random sampling from probability distributions, it is worth considering how computers can generate random numbers at all. As deterministic machines, computers cannot generate truly random numbers. Instead, they must rely on an algorithm which, given an initial value known as a *seed*, generates a sequence of numbers deterministically which nevertheless obey the statistical properties of being uniformly randomly distributed. Such an algorithm is known as a *pseudorandom number generator* (PRNG).

Poor PRNGs can be subject to a variety of problems, such as short periods before the sequence of numbers repeats, strong correlation between successive numbers in the sequence and a nonuniform distribution, particularly as the dimensionality increases (e.g., Press et al., 2007) A commonly used PRNG that does not suffer from these issues is the Mersenne Twister (Matsumoto and Nishimura, 1998), so named as its period is the Mersenne prime $2^{19937} - 1$.

Simple sampling methods

We now discuss several simple methods for sampling from probability distributions. The first of these is *inverse transform sampling* (e.g., Devroye, 1986). This method

utilises *cumulative distribution functions* (CDFs), an alternative way of specifying probability distributions, defined as

$$F(x) = \int_{-\infty}^{x} P(\theta)\, \mathrm{d}\theta \tag{3.35}$$

where $P(\theta)$ is the PDF represented by the CDF $F(x)$. The inverse cumulative distribution $F^{-1}(p)$ is then defined as the inverse of this function such that $F\left(F^{-1}(p)\right) = p$. Inverse transform sampling then proceeds by noting that, given uniformly distributed samples $u_i \sim U(0,1)$, the transformed samples $x_i = F^{-1}(u_i)$ are distributed according to the distribution $P(\theta)$ as desired.

Inverse transform sampling is an efficient sampling method when the inverse cumulative function $F^{-1}(p)$ can be evaluated. However, obtaining this function in general can be difficult, restricting inverse transform sampling to analytic distributions in practice.

An alternative sampling method suitable for distributions that are not known in closed form is *rejection sampling* (e.g., Gelman et al., 2013). This is a method of sampling from a distribution $P(\boldsymbol{\theta})$ by utilising another probability distribution $Q(\boldsymbol{\theta})$, known as the proposal distribution. This proposal distribution must be greater than the target distribution $Q(\boldsymbol{\theta}) > P(\boldsymbol{\theta})$ over the entire support of $\boldsymbol{\theta}$. However, neither of these distributions needs to be normalised to unity. Rejection sampling is therefore suitable for parameter inference problems where the normalising constant is not known.

The method proceeds as follows. A value $\boldsymbol{\theta}^i \sim Q(\boldsymbol{\theta})$ is sampled from the proposal distribution. It must therefore be possible to sample from the proposal distribution, e.g., by choosing a scaled uniform distribution, or by using inverse transform sampling. A uniform random number is also sampled $u \sim U(0,1)$. If $u < \frac{P(\boldsymbol{\theta}^i)}{Q(\boldsymbol{\theta}^i)}$, the sample is accepted. Otherwise, the sample is rejected.

While rejection sampling has the benefit of enabling sampling from unnormalised distributions, it has some downsides. The efficiency of the method relies on the choice of proposal distribution. A poor choice can lead to many samples being rejected, making the sampling inefficient. Adaptive rejection sampling methods (e.g., Gilks and Wild, 1992) can help with this by altering the proposal distribution as sampling proceeds to increase efficiency. Nevertheless, many samples can still be rejected, particularly in high dimensional parameter spaces, an example of the *curse of dimensionality*.

Markov chain Monte Carlo

This simple sampling methods discussed above are not generally suitable for Bayesian inference problems. We discuss here a widely used alternative, *Markov chain Monte Carlo* (MCMC). This is a family of sampling methods of which many variants exist, allowing a wide variety of otherwise difficult posterior distributions to be sampled. Below, we summarise the properties of MCMC following Bishop (2006).

MCMC produces a series of samples that form a *Markov chain* with an equilibrium distribution equal to the *target distribution*, the probability distribution to be sampled. Markov chains are a sequence of random variables $\boldsymbol{\theta}^i$ for $i \in \{1 \dots N\}$ where each value depends only on the previous value, i.e., the conditional distribution

$$P(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^1 \dots \boldsymbol{\theta}^{i-1}) = P(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}^{i-1}) \,. \tag{3.36}$$

This is known as the Markov property. MCMC then proceeds by making a series of steps through parameter space to form this chain, each step conditioned on the current position. These steps are controlled by the transition probability $T(\boldsymbol{\theta}^i \to \boldsymbol{\theta}^{i+1}) \equiv P(\boldsymbol{\theta}^{i+1} \mid \boldsymbol{\theta}^i)$.

For the marginal distribution $P(\boldsymbol{\theta}^i)$ to be stationary, the chain should obey the property known as *detailed balance*, where

$$P(\boldsymbol{\theta})T(\boldsymbol{\theta} \to \boldsymbol{\theta}') = P(\boldsymbol{\theta}')T(\boldsymbol{\theta}' \to \boldsymbol{\theta}) \,. \tag{3.37}$$

The Markov chain should also converge to the desired target distribution, irrespective of the starting point $\boldsymbol{\theta}^{i=1}$. A Markov chain that has this property is said to be *ergodic*.

MCMC produces a Markov chain that obeys both of these properties, allowing MCMC to be used to sample from a specified posterior distribution, as desired. However, the Markov property results in consecutive samples being correlated. If independent samples from the target density are required, a shorter chain can be created by selecting only every n points from the original chain, a process known as *thinning* (e.g., Gelman et al., 2013). In addition, while the ergodicity of the chain guarantees that it will eventually converge to the target distribution, the beginning of the chain can be unrepresentative if a poor starting point is chosen. A solution to this is to *burn in*, discarding the beginning of the chain and keeping only later samples. We discuss the details of several variants of MCMC below.

Metropolis and Metropolis-Hastings

The earliest MCMC variant is known as the *Metropolis algorithm* (Metropolis et al., 1953). In order to sample from the target distribution $P(\boldsymbol{\theta})$, this method makes steps from the current position $\boldsymbol{\theta}^i$ in parameter space by randomly sampling a new position $\boldsymbol{\theta}'$ from a proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^i)$ which is symmetric, i.e., $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^i) = q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}')$. This new position is then accepted with a probability given by

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}^i) = \min\left(1, \frac{P(\boldsymbol{\theta}')}{P(\boldsymbol{\theta}^i)}\right) . \tag{3.38}$$

This can be achieved by sampling a uniform random number $u \sim U(0, 1)$, and accepting the new sample if $u < A(\boldsymbol{\theta}^{i+1}, \boldsymbol{\theta}^i)$. Note that when the step has resulted in an increase in the target density $P(\boldsymbol{\theta}^{i+1}) > P(\boldsymbol{\theta}^i)$, the new sample is always accepted. If the new sample is accepted, the Markov chain is updated $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}'$. If the new sample is rejected, the current value of the parameter should be repeated as the new value, i.e., $\boldsymbol{\theta}^{i+1} = \boldsymbol{\theta}^i$.

This simple method can be generalised into the *Metropolis-Hastings* algorithm (Hastings, 1970) to allow a nonsymmetric proposal distribution $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^i)$. In this case, the acceptance probability is modified to be

$$A(\boldsymbol{\theta}', \boldsymbol{\theta}^i) = \min\left(1, \frac{P(\boldsymbol{\theta}')}{P(\boldsymbol{\theta}^i)} \frac{q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}')}{q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^i)}\right) . \tag{3.39}$$

The original Metropolis algorithm is now a special case of this more general form when $q(\boldsymbol{\theta}' \mid \boldsymbol{\theta}^i) = q(\boldsymbol{\theta}^i \mid \boldsymbol{\theta}')$. Note that the target density appears only as a ratio in the acceptance probability. As a result, these methods are able to sample from distributions where the normalising constant is not known, as is the case for many Bayesian parameter inference problems.

The efficiency of a Metropolis-Hastings sampler is dependent on the choice of proposal distribution which should be tuned for the specific target density. A common choice of proposal distribution is a Gaussian centred on the current position in parameter space.

Hamiltonian Monte Carlo

As the dimensionality of the target distribution increases, a Metropolis-Hastings sampler can struggle to explore the space efficiently, with the ratio of samples accepted to those rejected decreasing (e.g., Gelman et al., 2013). In parameter spaces with

more than ≈ 10 dimensions, the efficiency of Metropolis Hastings is extremely low. Given enough time, the samples will converge to the target distribution. In practice, however, this convergence requires too many samples to be feasible, and Metropolis Hastings cannot be used to sample to target distribution in a reasonable amount of computational time. Posterior distributions for problems of interest, particularly those derived from BHMs, can be higher dimensional than this, e.g., Jasche et al. (2015) constructs a posterior distribution of $> 10^7$ dimensions in order to infer large scale structure density fields from observed galaxy distributions. Metropolis Hastings is therefore not a suitable choice of sampler for these problems.

Hamiltonian Monte Carlo (HMC, Duane et al., 1987) is an MCMC variant that is able to sample these very high dimensional spaces. To do this, it utilises not only the value of the density but also its gradient, evolving a dynamical system through a phase space to explore parameter space efficiently. This phase space has twice the number of parameters as the parameter space, consisting of the *positions* given by the parameters $\boldsymbol{\theta}$, and the *momenta* $\boldsymbol{p}$. The evolution of these quantities is described by a Hamiltonian function

$$H(\boldsymbol{\theta}, \boldsymbol{p}) = U(\boldsymbol{\theta}) + K(\boldsymbol{p}) \,, \tag{3.40}$$

where the *potential energy* $U(\boldsymbol{\theta})$ is the negative log of the target density, i.e.,

$$U(\boldsymbol{\theta}) = -\log P(\boldsymbol{\theta}) \,, \tag{3.41}$$

and the *kinetic energy* $K(\boldsymbol{p})$ is given by

$$K(\boldsymbol{p}) = \frac{1}{2} \boldsymbol{p}^T \underline{\boldsymbol{M}}^{-1} \boldsymbol{p} \,. \tag{3.42}$$

The matrix $\underline{\boldsymbol{M}}$ is known as the *mass matrix* and provides a way to tune the performance of the sampling algorithm. It is possible to generalise this procedure so that the mass matrix varies in parameter space, i.e., $\underline{\boldsymbol{M}} \to \underline{\boldsymbol{M}}(\boldsymbol{\theta})$. This generalisation is referred to as *Riemannian Hamiltonian Monte Carlo* (Girolami and Calderhead, 2011), and can aid exploration in highly correlated parameter spaces.

One iteration of HMC sampling then proceeds as follows. First, an initial momentum $\boldsymbol{p}_0$ is sampled from a Gaussian distribution with mean of zero and covariance given by the mass matrix $\underline{\boldsymbol{M}}$. Next, the position and momentum are evolved according to Hamilton's equations, given by (e.g., Neal, 2012)

$$\frac{\mathrm{d}\theta_j}{\mathrm{d}t} = (\underline{\boldsymbol{M}}^{-1} \boldsymbol{p})_j \tag{3.43}$$

and

$$\frac{\mathrm{d}p_j}{\mathrm{d}t} = -\frac{\partial U}{\partial \theta_j} \tag{3.44}$$

for dimension j. To solve this in practice, a numerical method such as *leapfrog integration* must be used. After N timesteps of the numerical integration, the resulting position and momentum vectors $\{\boldsymbol{\theta}', \boldsymbol{p}'\}$ are taken to be the new proposal values. The acceptance probability is then given by

$$A(\{\boldsymbol{\theta}', \boldsymbol{p}'\}, \{\boldsymbol{\theta}^i, \boldsymbol{p}_0\}) = \min\left(1, \mathrm{e}^{-(H(\boldsymbol{\theta}', \boldsymbol{p}') - H(\boldsymbol{\theta}^i, \boldsymbol{p}_0))}\right), \tag{3.45}$$

where $\boldsymbol{\theta}^i$ is the previous parameter position, and $\boldsymbol{p}_0$ is the sampled initial momentum. Note that, if it were not for errors introduced by the numerical integration, the Hamiltonian would be conserved and every sample would be accepted. As in Metropolis-Hastings, if the proposed sample is rejected, the previous parameter value is repeated in it place. Also note that the momentum vector is always discarded, as a new momentum vector is sampled at the beginning of each iteration.

The efficiency of HMC is sensitive to the number of timesteps performed in the numerical integration step. The No-U-Turn Sampler (NUTS, Hoffman and Gelman, 2011) is a method to adaptively set this number of timesteps. This allows HMC to be easily applied to a variety of problems without hand-tuning using probabilistic inference software packages (e.g., Salvatier et al., 2016; Carpenter et al., 2017). This application is also aided by the automatic differentiation methods described in section 3.2.3.

Gibbs sampling

The MCMC methods discussed above all rely on being able to evaluate the complete joint distribution for the target density. For Bayesian inference problems, this is the posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{d})$. Instead, *Gibbs sampling* (Geman and Geman, 1984) is an MCMC method that samples from a joint distribution specified only by its conditional distributions.

Consider a D-dimensional joint posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{d})$. The conditional distributions of each of its parameters are then given by $P(\theta_j \mid \theta_1 \ldots \theta_{j-1}, \theta_{j+1} \ldots \theta_D, \boldsymbol{d})$ for dimension $j \in \{1 \ldots D\}$. Iteration i of Gibbs sampling then proceeds by sampling $\theta_j^i \sim P(\theta_j \mid \theta_1 \ldots \theta_{j-1}, \theta_{j+1} \ldots \theta_D, \boldsymbol{d})$ for each dimension j. Each of these samples is accepted, forming the new parameter vector $\boldsymbol{\theta}^i$.

Gibbs sampling can be a useful sampling method when the target density is most easily written in terms of its conditional distributions, as is often the case in hierarchical models (e.g., Alsing et al., 2017). In fact, these conditional distributions can

be obtained directly from the DAG of a model, a graphical representation commonly used to represent hierarchical models as described in section 3.1.8. The samples from each conditional can be obtained through any other sampling method. Gibbs sampling is a particularly efficient choice when the conditional distributions can be analytically sampled. However, MCMC methods such as Metropolis-Hastings and HMC can also be used, forming Metropolis-within-Gibbs (e.g., Liu and Tong, 2019) and HMC-within-Gibbs (e.g., Dang et al., 2017) samplers respectively.

Nested sampling

As discussed in section 3.1.6, evaluating the multidimensional integral required to calculate the evidence $P(\boldsymbol{d} \mid \mathcal{M})$ can be computationally challenging. Nested sampling (Skilling, 2006) is a Monte Carlo sampling method that samples the posterior, efficiently calculating the evidence. The resulting samples, appropriately weighted, can also be used for parameter estimation.

Nested sampling reduces the computational burden of calculating the evidence by transforming the multidimensional integral into one over a single dimension only. Skilling (2006) first defines the prior volume element $\mathrm{d}X = \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}$ and the function

$$X(\lambda) = \int_{\mathcal{L}(\boldsymbol{\theta};\,\boldsymbol{d})>\lambda} \pi(\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta}\,, \tag{3.46}$$

where $\mathcal{L}(\boldsymbol{\theta};\,\boldsymbol{d}) \equiv P(\boldsymbol{d} \mid \boldsymbol{\theta}, \mathcal{M})$ is the likelihood, $\pi(\boldsymbol{\theta} \mid \mathcal{M}) \equiv P(\boldsymbol{\theta})$ is the prior and $\boldsymbol{\theta}$ is the N-dimensional parameter vector. This function $X(\lambda)$ gives the volume of the prior where the likelihood is greater than the threshold λ. This volume is therefore contained within a boundary in parameter space, along which the likelihood is constant. This boundary is known as an *isolikelihood contour*. The evidence can then be written as the one-dimensional integral

$$\mathcal{Z} = \int_0^1 L(X)\,\mathrm{d}X\,, \tag{3.47}$$

where the function $L(X)$ is the inverse of equation 3.46, i.e., $L(X)$ is the value of the likelihood corresponding to the isolikelihood contour that contains X prior volume. The problem of estimating the evidence then reduces to sampling a series of these likelihoods $L(X_i)$ where $X_{i+1} < X_i$, i.e., progressively smaller prior volumes nested within one another. Equation 3.47 can then be calculated using a simple quadrature integration method.

This sampling process starts by initialising a pool of n random points sampled from the prior. These points, referred to as the *live points*, each correspond to a

particular likelihood value, and so each define an isolikelihood contour. The key insight of Skilling (2006) is that, since the live points were sampled from the prior, the prior volumes X contained within these contours are therefore uniformly sampled $X \sim U(0, 1)$.

The live point with the lowest likelihood is then removed as the first sample, and the prior volume contained within its isolikelihood contour labelled X_0. This point is then replaced with a new point sampled from the prior, but with the condition that its likelihood $L > L(X_0)$; that is, it is contained within the isolikelihood contour of the first sample. This therefore corresponds to uniformly sampling the prior volume $X \sim U(X_0, 1)$. This procedure defines a single iteration of the nested sampling algorithm which repeats with the new set of live points.

At iteration i, the prior volumes of the isolikelihood contours of all the live points are sampled uniformly between 0 and X_{i1}. The lowest likelihood point is then, by definition, the point that defines the contour containing the largest prior volume X_i. Thus, the prior volume X_i can be estimated statistically using the distribution of the N^{th} largest value in a uniform sample of N points. Values of this sort are known as *order statistics* (e.g., David and Nagaraja, 2006). The prior volume at iteration i can then be approximated as $X_i \approx \mathrm{e}^{i/n}$.

This iterative process continues until a precision threshold for the evidence is reached. Once it has, the m removed points are combined with the n remaining points and used to approximate the evidence as (Feroz et al., 2009)

$$\mathcal{Z} = \sum_{j=1}^{n+m} L(X_j) w_j \,, \tag{3.48}$$

where $w_j = \frac{X_{j-1} - X_{j+1}}{2}$ for a point removed during iteration j and $w_j = \frac{X_m}{n}$ for the remaining live points. In addition, this combination of points can be considered to be weighted samples from the posterior that can be used for parameter estimation. The weighting of point j is given by

$$p_j = \frac{L(X_j) w_j}{\mathcal{Z}} \,, \tag{3.49}$$

corresponding to the fractional contribution of point j to the total evidence approximation.

Several variations of the nested sampling method exist (e.g., Brewer et al., 2009; Feroz et al., 2009; Handley et al., 2015; Higson et al., 2017). Sampling from within an isolikelihood contour is the most computationally expensive part of the nested sam-

pling procedure. The work presented in chapter 6 uses the MultiNest nested sampling method (Feroz et al., 2009), a method for improving the efficiency of this step by approximating the contour with a series of ellipses; this method is explained in more detail in section 6.3.2.

3.2 Machine Learning

The Bayesian methods discussed in section 3.1 all involve the assumption of a physical forward model, typically encoded in the likelihood function. In contrast, *machine learning* methods make no reference to a physical model. Instead, these methods fit extremely flexible models to large datasets known as *training sets*, using the information obtained to make predictions. An advantage of these methods is the ease with which they scale to very large datasets, a property that will become increasingly important as the size of datasets grows, as discussed above. This section discusses machine learning methods in general and their place within a scientific analysis.

3.2.1 Inference vs prediction

The applications of a statistical or machine learning data analysis can broadly be summarised into two types. Following the notation of James et al. (2013), we refer to these as *inference* and *prediction*.

Bayesian statistical analyses are typically used to perform inference. Either a particular data generating process is assumed, as is the case for parameter inference, or a series of such processes are, as in model comparison. When applied to the analysis of cosmological data, this data generating process is typically informed by a theoretical model that, as best as possible, approximates the true process of data generation found in nature. Inference is therefore concerned with drawing conclusions about this model, and by extension, the natural system under consideration. These conclusions can take the form of posterior distributions over model parameters and relative probabilities of competing models.

The application of statistical inference to scientific problems is therefore clear; making inferences about processes found in nature is the *goal* of the scientific method. In contrast, the goal of a predictive analysis is not to understand the true data generating process at all, simply to predict the effects of it; given a particular input, what can be predicted about the output? Machine learning methods are typically used to preform a predictive analysis. In these cases, the function that maps the inputs to

their respective outputs can be a black box, lacking interpretability in any practical sense (e.g., Lipton, 2016).

This lack of interpretability of machine learning models has been noted (e.g., Rudin, 2018) as a potential problem when these methods are applied to problems where a full understanding of the inference is important. An example of a problem arising from a lack of interpretability is the existence of adversarial examples (Goodfellow et al., 2014), data that have been modified in a minimal way to cause the predictions of machine learning methods to lose considerable accuracy. These modifications can be extremely small, such as the addition of noise to an image (Kurakin et al., 2016), resulting in changes that cannot be detected by a human but nevertheless cause a machine learning method to misclassify the image. These examples highlight the difficulty of interpreting the workings of a highly complex black box method.

Given these concerns, it is reasonable to ask what the role of machine learning methods in a scientific analysis should be. It is the view of the author that the distinction between inference and prediction can guide this. While it is possible for a machine learning method to be trained to predict model parameters from simulated data to a high accuracy (e.g., Ravanbakhsh et al., 2017), this task is one of inference and would therefore benefit from interpretability. Model parameters are only meaningful quantities in the context of a model, and for the final conclusions of a scientific analysis to be believed, they should be explainable.

However, prediction tasks are also a common part of cosmological data analysis pipelines. The problem of estimating redshifts from photometric data, the subject of this thesis, is an example of this. Indeed, photometric redshifts are an area that has seen a proliferation of machine learning-based approaches; a variety of these are discussed in more detail in section 4.1.2. The complete data generating process for photometric redshift problems is complex, potentially encompassing star formation and the details of the chemistry involved therein. For the purpose of cosmological data analysis, this level of detail is both unnecessary and infeasible. In addition, photometric redshifts are necessarily a noisy estimate with catastrophic failures due to the loss of information when integrating a spectrum over photometric filters. The pragmatic concern of photometric redshift methods is therefore the accuracy of their predictions on a representative set of data, a goal which aligns with that of machine learning methods.

3.2.2 Supervised and unsupervised learning

Various different machine learning methods for data analysis tasks exist. However, these methods can be characterised into two distinct types by the form of the data on which they are trained. These are known as *supervised* and *unsupervised* learning[4].

Supervised learning involves methods that use training data that contains both an input $\boldsymbol{x}$ and output $\boldsymbol{y}$. Such data is sometimes referred to as *labelled data*. The aim of the machine learning method is then to learn a function which maps from the inputs to the outputs. This can then be used to predict the output $F(\boldsymbol{x}) = \hat{y}$ corresponding to the inputs of other previously unseen data, known as *test data*. To do this, a user must specify a *loss function* $L(y, \hat{y})$, a function which quantifies the accuracy of the prediction as a scalar. Supervised learning is therefore an optimisation problem (e.g., Bishop, 2006) where the loss function is minimised, given a particular training set. These methods typically employ a gradient-based optimisation method, as described in section 3.2.3. Examples of supervised learning methods include random forests (Breiman, 2001), Gaussian processes (Rasmussen and Williams, 2005) and neural networks (e.g., LeCun et al., 2012).

In contrast to supervised learning, unsupervised learning methods utilise *unlabelled data*, where the training set has no output to be predicted. Instead, these methods aim to find unspecified patterns in the data. An example application of this is clustering, where samples from a dataset are categorised into several clusters based on their similarity. K-means clustering (e.g., MacQueen, 1967; Jain, 2010) is an example of a clustering algorithm which separates each sample into a single cluster. A probabilistic generalisation of this is known as Gaussian mixture models (GMMs) (e.g., Fraley and Raftery, 1998), where each cluster is comprised of a multivariate Gaussian, and each sample is probabilistically assigned to clusters. GMMs are also an example of another application of unsupervised learning known as *density estimation*, where samples in the training set are used to estimate the data-space probability distribution from which they were sampled. These applications of unsupervised learning can be used for several purposes, including the preprocessing of data for supervised learning methods, known as *feature engineering* or *feature extraction* (e.g., Hastie et al., 2005), or for deriving the conditional probability distributions; see chapter 7 for an example of using conditional distributions derived from GMMs to infer photometric redshifts.

[4]A third type of machine learning method also exists, known as *reinforcement learning* (e.g., Li, 2017). This is a method to train software-controlled *agents* to make decisions by rewarding them based on completing desirable goals. These methods are not typically applied in data analysis contexts, and as such, we do not discuss them further here.

3.2.3 Gradient-based optimisation

Optimisation is the numerical problem of maximising the value of an *objective function* by modifying the values of the parameters that control it. In order to efficiently optimise objective functions of many parameters, it is essential to use an optimisation method that uses the derivative of the function, known as *gradient-based optimisation* methods. This section reviews a variety of these methods.

Automatic differentiation

In order to optimise functions using their gradient, that gradient must, of course, be specified in some way. While the gradients for a particular model can be derived analytically for evaluation, this is a time-inefficient process, particularly as the complexity of the model and the number of its parameters increase. An alternative to this process is *automatic differentiation* (e.g., Baydin et al., 2018), a method by which gradients can be calculated for a particular model automatically.

Unlike numerical methods such as finite differencing, gradients calculated through automatic differentiation are not approximations, but are exact. Mathematical models $M(x) = y$ are specified in terms of primitive operations $f(x)$ such as addition, multiplication and exponentiation, i.e., $M(x) \equiv h(g(f(x)))$. The derivative of each of these primitive operations with respect to their input parameters $\nabla f(x)$ is known. As the model is evaluated, the application of each of these operations is recorded by the automatic differentiation software. The final derivative of the output value with respect to the inputs can then be calculated using the chain rule.

Due to the high complexity of machine learning models that require gradients for optimisation, automatic differentiation has been implemented in several machine learning software packages (e.g., Maclaurin et al., 2015; Abadi et al., 2016; Paszke et al., 2017). While the gradient-based optimisation methods discussed in the section are a common use of gradients within statistical inference and machine learning applications, another notable example is Hamiltonian Monte Carlo, discussed in section 3.1.9. Many implementations of HMC also therefore make use of automatic differentiation (e.g., Carpenter et al., 2015).

Stochastic gradient descent

The simplest gradient-based optimisation method is known as *gradient descent*. At each iteration, this method simply steps through parameter space in the direction

of the negative gradient, i.e., *downhill*. Consider a multivariate function $F(\boldsymbol{\theta})$. At iteration i, the gradient of the function $\nabla F(\boldsymbol{\theta}_i)$ is evaluated at the current position in parameter space $\boldsymbol{\theta}$. The parameters are then updated to be

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i - \boldsymbol{u}_i \,, \tag{3.50}$$

where

$$\boldsymbol{u}_i = \eta \nabla F(\boldsymbol{\theta}_i) \tag{3.51}$$

is the update at iteration i. The step size η is a hyperparameter controlling the size of the move at each iteration. This process continues until the optimisation is deemed to have converged, typically measured through the change in the value of the function between iterations, i.e., $|F(\boldsymbol{\theta}_i) - F(\boldsymbol{\theta}_{i-1})|$.

While gradient descent is simple to implement, exactly evaluating the function $F(\boldsymbol{\theta})$ can be costly when the model is complex or the dataset is large. A faster alternative to this is *stochastic gradient descent* (SGD, e.g., Bottou, 2010). A common scenario in machine learning optimisation problems is that the objective function $F(\boldsymbol{\theta} \mid \{\boldsymbol{d}\})$ involves a summation over samples in the dataset, i.e.,

$$F(\boldsymbol{\theta} \mid \{\boldsymbol{d}\}) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{\theta} \mid \boldsymbol{d}_i) \,, \tag{3.52}$$

where i indexes the sample in a dataset of N samples. Stochastic gradient descent replaces this gradient with an approximation given by the gradient evaluated on a single sample j from the dataset, i.e.,

$$\nabla F(\boldsymbol{\theta} \mid \{\boldsymbol{d}\}) \approx \nabla f(\boldsymbol{\theta} \mid \boldsymbol{d}_j) \,. \tag{3.53}$$

It can be seen that this is an unbiased estimate of the gradient by taking the expectation over j, i.e.,

$$\mathrm{E}_i \left[\nabla f(\boldsymbol{\theta} \mid \boldsymbol{d}_j)\right] \equiv \frac{1}{N} \sum_{j=1}^{N} \nabla f(\boldsymbol{\theta} \mid \boldsymbol{d}_j) = \nabla F(\boldsymbol{\theta} \mid \{\boldsymbol{d}\}) \,. \tag{3.54}$$

This noisy estimate of the gradient is then used in place of the exact gradient calculation in equation 3.50. This procedure can be generalised to use small sets of samples randomly selected from the full dataset in order to increase the accuracy of the gradient approximation. This is known as *mini-batch stochastic gradient descent*. SGD and its mini-batch variant have found widespread use within machine learning applications (see, e.g., LeCun et al., 2012).

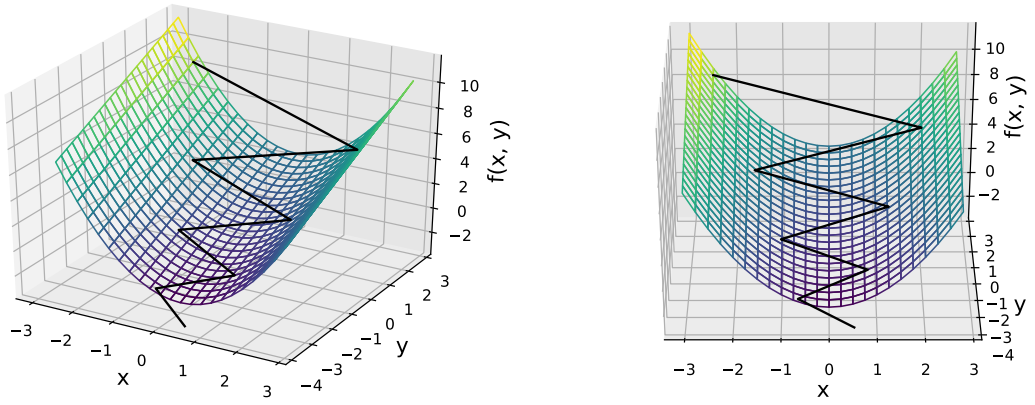

Figure 3.2: An example of the oscillatory behaviour that can occur when optimising a valley-shaped function, e.g., $f(x, y) = x^2 + y$ using gradient descent.

Improvements to gradient descent

Several variants of SGD have been proposed which improve its ability to optimise objective functions of many parameters. We summarise a review of these in Ruder (2016) here. Firstly, the assumption of a single step size η in SGD is often sub-optimal. Optimisation algorithms that utilise *adaptive step sizes* change this by modifying the step size separately for each dimension. This was first introduced by the AdaGrad algorithm (Duchi et al., 2011).

Another improvement to SGD is known as *momentum* (Rumelhart et al., 1986). A common problem for gradient descent methods occurs in areas of the parameter space where the gradient in one direction is significantly greater than in another, i.e., a *valley*. These can cause the optimiser to oscillate and therefore progress through the space slowly, as shown in Figure 3.2. Momentum provides a simple way to counter this issue by modifying the update at iteration i to be a linear combination of the gradient step and the previous update, i.e.,

$$\boldsymbol{u}_i = \eta \nabla F(\boldsymbol{\theta}_i) + \alpha \boldsymbol{u}_{i-1} \,. \tag{3.55}$$

By doing this, movement in a consistent direction is amplified as the updates accumulate, while oscillatory motion is damped as the updates cancel. The Adam optimiser (Kingma and Ba, 2014) is an example of a method that incorporates momentum, in addition to adaptive step sizes.

Finally, *Nesterov accelerated gradients* (Nesterov, 1983) are a modification to the momentum method above that improves convergence by evaluating the gradient step

after making the momentum step, i.e., defining the momentum step to be

$$\boldsymbol{m}_i = \alpha \boldsymbol{u}_{i-1} \,, \tag{3.56}$$

the update at iteration i is given by

$$\boldsymbol{u}_i = \eta \nabla F(\boldsymbol{\theta}_i - \boldsymbol{m}_i) + \alpha \boldsymbol{u}_{i-1} \,. \tag{3.57}$$

This modification has been incorporated into the Adam optimiser (Dozat, 2016).

3.2.4 Variational inference

While machine learning methods have typically been formulated initially as deterministic function approximators, they are increasingly being interpreted in a probabilistic manner (e.g., Murphy, 2012). For example, neural networks can be put into a framework of Bayesian inference by inferring posterior distributions over their weight parameters (e.g., Graves, 2011; Neal, 2012; Blundell et al., 2015). A method of performing probabilistic inference on these models and other Bayesian models in big data settings is *variational inference* (VI).

The methods described in section 3.1.9 aim to draw samples from a posterior distribution in order to make inferences. By expending more computational effort to increase the number of samples, derived quantities such as expectation values converge to their correct values. Sampling is therefore an example of an *exact inference* method. However, as discussed above, the volume of data available for cosmological analysis is set to increase significantly. Exact inference methods such as these can be computationally expensive, and thus can be difficult to scale to very large datasets.

An alternative approach to scaling Bayesian methods to complex problems and large datasets is to use *approximate inference*. These methods only approximate the results an exact inference would achieve, but in return, are able to perform inference on problems that are inaccessible through exact methods in practice. Likelihood-free methods (e.g., Turner and Zandt, 2012; Leclercq, 2018; Alsing et al., 2019; Taylor et al., 2019) perform inference on models where a likelihood cannot be evaluated. Instead, the forward model is specified, and data is simulated from the model conditional on its parameters.

Variational inference is another example of approximate inference. By recasting posterior inference into an optimisation problem, VI is able to scale to much larger datasets than is practically possible through sampling methods. The optimisation

problem that VI solves is approximating the target posterior distribution $P(\boldsymbol{\theta} \mid \boldsymbol{d})$ with another, simpler distribution, known as the *variational distribution* $q(\boldsymbol{\theta} \mid \boldsymbol{\phi})$, parametrised by $\boldsymbol{\phi}$. This is done by minimising the KL Divergence (Kullback and Leibler, 1951) introduced in equation 3.25 between the two distributions, i.e., the parameters of the variational distribution are given by

$$\boldsymbol{\phi}^* = \underset{\boldsymbol{\phi}}{\operatorname{argmin}}\, D_{\mathrm{KL}}\big(q(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \mid P(\boldsymbol{\theta} \mid \boldsymbol{d})\big)\,. \tag{3.58}$$

Evidence lower bound

We now summarise how VI solves this optimisation problem, following Blei et al. (2017). We first note that the KL divergence defined in equation 3.25 can be rewritten in terms of expectations by using the property that $\mathrm{E}\left[X + Y\right] = \mathrm{E}\left[X\right] + \mathrm{E}\left[Y\right]$, giving

$$D_{\mathrm{KL}}\big(q(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \mid P(\boldsymbol{\theta} \mid \boldsymbol{d})\big) = \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\right] - \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\log P(\boldsymbol{\theta} \mid \boldsymbol{d})\right]\,, \tag{3.59}$$

where $\mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\dots\right] \equiv \int q(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \dots \,\mathrm{d}\boldsymbol{\theta}$ indicates an expectation over $q(\boldsymbol{\theta} \mid \boldsymbol{\phi})$. Since the KL divergence requires normalised distributions, this cannot be evaluated without calculating the expensive evidence integral as described in section 3.1.6. Hence, inserting $P(\boldsymbol{\theta} \mid \boldsymbol{d}) = P(\boldsymbol{\theta}, \boldsymbol{d})/P(\boldsymbol{d})$, equation 3.59 becomes

$$\begin{aligned} D_{\mathrm{KL}}\big(q(\boldsymbol{\theta} \mid \boldsymbol{\phi}) \mid P(\boldsymbol{\theta} \mid \boldsymbol{d})\big) &= \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\right] - \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\log P(\boldsymbol{\theta}, \boldsymbol{d})\right] + \log P(\boldsymbol{d}) \\ &\equiv -\mathrm{ELBO}(\boldsymbol{\phi}) + \log P(\boldsymbol{d})\,, \end{aligned} \tag{3.60}$$

where the evidence has been moved outside of the expectation as it is constant w.r.t. $\boldsymbol{\theta}$. In order to avoid calculating the evidence, VI instead optimises an alternative objective function known as the evidence lower bound $\mathrm{ELBO}(\boldsymbol{\phi})$, given by the negative of the first two terms of equation 3.60, i.e.,

$$\mathrm{ELBO}(\boldsymbol{\phi}) = \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\log P(\boldsymbol{\theta}, \boldsymbol{d}) - \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\right]\,, \tag{3.61}$$

Since the evidence $P(\boldsymbol{d})$ is constant, maximising the evidence lower bound is equivalent to minimising the KL-divergence, as desired. This function can now be optimised in order to approximate the posterior. We discuss two possible methods for performing this optimisation below.

Mean-field approximation

Early applications of variational inference (e.g., Attias, 1999; Ghahramani and Beal, 2001; Blei et al., 2003, 2006) derived analytic update rules to the variational distributions in order to optimise them. To do this, they used a particular form of variational distribution known as the *mean-field approximation*, where each variational parameter is assumed independent, i.e.,

$$q(\boldsymbol{\theta} \mid \boldsymbol{\phi}) = \prod_{j=1}^{D} q_j(\theta_j \mid \phi_j)\,, \tag{3.62}$$

for a D-dimensional parameter vector. The optimal variational distribution for dimension j, holding other dimensions constant, can then be found to be (e.g., Bishop, 2006)

$$\log q_j^*(\theta_j \mid \phi_j) \propto \mathrm{E}_{\prod_{j' \neq j} q_{j'}(\theta_{j'} \mid \phi_{j'})} \left[\log P(\theta_j \mid \theta_1 \ldots \theta_{j-1}, \theta_{j+1} \ldots \theta_D, \boldsymbol{d})\right]\,, \tag{3.63}$$

where $\mathrm{E}_{\prod_{j' \neq j} q_j(\theta_{j'} \mid \phi_{j'})}[\ldots] \equiv \int \ldots \prod_{j' \neq j} q_{j'}(\theta_{j'} \mid \phi_{j'})\ \mathrm{d}\theta_{j'}$ indicates an expectation over the product of variational distributions excluding dimension j. For posteriors where the conditional distributions $P(\theta_j \mid \theta_1 \ldots \theta_{j-1}, \theta_{j+1} \ldots \theta_D, \boldsymbol{d})$ are in the exponential family of distributions[5] and the priors on all parameters are conjugate as described in section 3.1.7, these expectations can be computed analytically to give variational distributions that are also in the exponential family.

Black box variational inference

In order to enable the more general application of variational inference, numerical methods must be employed. *Black box variational inference* (Ranganath et al., 2013) is a method to optimise the variational parameters using stochastic gradient descent, as described in section 3.2.3. This allows VI to be applied to a wider variety of models than the analytic method described above.

To optimise the variational parameters in this way, we require the gradient of the evidence lower bound defined in equation 3.61, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}} \mathrm{ELBO}(\boldsymbol{\phi}) = \frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}} \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})} \left[\log P(\boldsymbol{\theta}, \boldsymbol{d}) - \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\right]\,. \tag{3.64}$$

[5] The exponential family of distributions contains many common distributions including the normal, beta, gamma and Dirichlet distributions. Likelihoods that belong to this family of distributions will always have a conjugate prior.

Ranganath et al. (2013) show that this gradient of the expectation can be rewritten as an expectation of a gradient, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}}\mathrm{ELBO}(\boldsymbol{\phi}) = \mathrm{E}_{q(\boldsymbol{\theta} \mid \boldsymbol{\phi})}\left[\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}} \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\big(\log P(\boldsymbol{\theta}, \boldsymbol{d}) - \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\big)\right] . \tag{3.65}$$

This expectation can then be approximated using samples from the variational distribution as

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}}\mathrm{ELBO}(\boldsymbol{\phi}) \approx \frac{1}{S}\sum_{s=1}^{S}\left[\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\phi}} \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\big(\log P(\boldsymbol{\theta}, \boldsymbol{d}) - \log q(\boldsymbol{\theta} \mid \boldsymbol{\phi})\big)\right] , \tag{3.66}$$

where $\boldsymbol{\theta}_i \sim q(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ for $s \in \{1 \dots S\}$. Equation 3.66 defines a noisy Monte Carlo approximation to the gradient as required by stochastic gradient descent as described in section 3.2.3. The requirements for applying VI have therefore reduced to being able to sample from the variational distribution $q(\boldsymbol{\theta} \mid \boldsymbol{\phi})$ and being able to evaluate the gradient inside the square brackets. These gradients can be obtained using automatic differentiation methods as described in section 3.2.3, a method known as *automatic differentiation variational inference* (ADVI, Kucukelbir et al., 2016).

Chapter 4

Photometric Redshifts

As discussed in section 2.3, obtaining the redshifts of observed sources is an integral part of utilising cosmological galaxy surveys. There are two distinct uses for these redshifts. Firstly, cosmological galaxy surveys typically separate their sources into several redshift bins, a process known as tomography. Doing this enables redshift-dependent measurements to be made without the complications of a full three-dimensional analysis, increasing the constraining power of the survey. Secondly, as detailed in section 2.3.3, theoretical predictions of angular power spectra that can be measured from cosmological galaxy surveys and used to constrain cosmology require knowing the redshift distribution of sources within each tomographic bin.

Spectroscopic observations allow the highest-precision determination of redshifts. By observing the spectrum of each source of interest, emission or absorption lines arising due to elements present in the source can be identified. Since these lines are at known rest-frame wavelengths, their observed wavelengths provides a way to measure the redshift of the source. In practice, this measurement is done by cross-correlating the spectrum that is observed with an *a priori* specified set of templates (e.g., Tonry and Davis, 1979; Baldry et al., 2014) or modelling the line profiles (e.g., Mink and Wyatt, 1995). This is discussed in more detail in section 2.3.1.

The downside to utilising spectroscopic redshift is that observing spectra in this way requires a large amount of telescope time. This therefore places a limit on the total number of sources that can have their redshifts spectroscopically determined to a sufficient signal-to-noise using a telescope with a particular sensitivity, given a finite amount of observation time.

One factor controlling the precision of cosmological constraints obtained from cosmological galaxy surveys utilising measurements such as cosmic shear and galaxy-galaxy lensing is the number density of sources observed. These lensing effects can only

be measured from galaxy shapes statistically since the intrinsic shapes of galaxies are unknown. These statistical measurements therefore rely on having a sufficiently high number density of galaxies to counter the statistical noise of this effect. Thus, in order to obtain high-precision constraints from cosmological galaxy surveys, observations of a large number of galaxies are required. As described above, this can be a problem for spectroscopic observations.

An alternative to spectroscopic redshifts is *photometric redshifts*. In contrast to spectroscopic observations which finely disperse the collected light, photometric observations generally utilise a small set of filters, imaging the sky in each filter. The flux of every source contained in these images can then be measured by summing the flux of each pixel associated with a source and subtracting the background flux level, referred to as *photometry*. A common tool to automatically identify the relevant pixels for each source is SExtractor (Bertin and Arnouts, 1996); see chapter 5 for a more detailed description of this method and a discussion about the complications to this process posed by the blending of sources. The result of this process is a vector of fluxes and associated uncertainties for each source of interest. Photometric redshift methods are then statistical methods which utilise these outputs in order to determine the redshift of the source.

By measuring the flux integrated over a series of colour filters centred at several wavelengths, photometry can be seen as providing a low-resolution analogue to galaxy spectra. The constraining power from these measurements typically comes from strong features of the underlying spectrum such as the Lyman and Balmer breaks (Salvato et al., 2019), absorption features at rest-frame wavelengths of 912Å and 3650Å respectively. The ability for photometric redshifts to determine the redshift of sources accurately therefore depends on the filters of the respective galaxy survey covering these redshifted features. Choosing these filters carefully for the population of sources of interest can significantly increase the precision of the resulting redshift estimates (Benítez et al., 2009).

The statistical nature of photometric redshifts is important for a key part of precision cosmology; namely, an accurate understanding of uncertainties in parameter constraints. To enable this, uncertainties arising from each step of the analysis should be accounted for and propagated onwards. In cosmological analyses, this is typically accomplished using a Bayesian framework (e.g., Hildebrandt et al., 2017; Troxel et al., 2017), allowing these uncertainties to be combined and marginalised over for the final constraints. It is therefore essential that photometric redshift methods provide not only point estimates of redshifts, but also a measure of their uncertainties.

The uncertainty associated with a redshift estimate are sometimes represented by a single number, i.e., a point estimate with an error bar. However, doing this necessitates making an assumption about how the error is distributed. Uncertainties in photometric redshifts can be highly non-Gaussian (e.g., Quadri and Williams, 2010), and so are poorly described by a single number such as the variance. Photometric redshift methods that instead characterise their results using a probability distribution function (PDF) can capture all of this information. PDFs are discussed in more detail in section 3.1.2.

Photometric redshifts can also suffer from degeneracies that result in high-redshift galaxies having similar colours to those at low redshifts (e.g., Graham et al., 2018). As a result, several well-separated redshifts are plausible, and an accurate representation of the uncertainty should reflect this. While this can be easily described with a multi-modal PDF, a single number can be misleading. Error bars that cover the full range of parameter space between the low- and high-redshift estimates do not show that redshifts between these are disfavoured, inflating uncertainties. Several photometric redshift methods are able to produce PDFs as their result. These are discussed in more detail in the following sections.

Ensuring that photometric redshifts are sufficiently accurate and precise is necessary for obtaining unbiased constraints on cosmological parameters. Huterer et al. (2006) found that future tomographic surveys would require the mean of each redshift bin to be known to a precision of 0.003, though this requirement can be reduced by self-calibration (e.g., Huterer et al., 2006; Sun et al., 2015; Samuroff et al., 2017) and combining weak lensing data with other cosmological probes such as baryonic acoustic oscillations (e.g., LSST Science Collaboration et al., 2009). Photometric redshifts are also important in the calibration of other systematics. Multiplicative biases in the measurement of shear can be detected and corrected for, provided that photometric redshifts of galaxies in the sample are unbiased (Hoekstra et al., 2017). Weak lensing shape measurement biases can themselves also be redshift dependent; without unbiased redshift estimates to make corrections, these can lead to biases of a few percent in the cosmological parameters σ_8 and w_0 (Semboloni et al., 2009).

A variety of photometric redshift approaches were compared by Schmidt et al. (2020) in the context of them being applied to future LSST (Ivezić et al., 2019) observations. To do this, they used a simulated galaxy catalogue derived from N-body simulations populated with galaxies using linear combinations of SDSS (Stoughton et al., 2002) spectra. The primary aim of this work was then to evaluate the metrics on which photometric redshift methods may be compared. This was done by proposing a photometric redshift method where the PDF of the redshift of each galaxy was

given by the redshift distribution of the simulated catalogue, irrespective of the input photometry. Such a method would be of little use in real applications, and therefore represents a useful test of metrics themselves; a metric on which this method scores highly is unlikely to informative. Nevertheless, Schmidt et al. (2020) found that only a single metric scored this method poorly, the conditional density estimate (CDE). The CDE can be estimated for an inferred redshift PDF $\hat{f}(z \mid \boldsymbol{X})$ conditioned on data $\boldsymbol{X}$ by

$$\mathrm{CDE} = \mathrm{E}_{\boldsymbol{X}}\left[\int \hat{f}(z \mid \boldsymbol{X})^2 \, \mathrm{d}z\right] - 2\mathrm{E}_{\boldsymbol{X},z}\left[\hat{f}(z \mid \boldsymbol{X})\right] + K\,, \tag{4.1}$$

where the first expectation is over the data, the second is over the data and redshift, and K is a constant. Schmidt et al. (2020) conclude by emphasising that an evaluation metric should be chosen to reflect scientific goals in order to be of most use.

Photometric redshifts are the subject of the research work of this thesis, where we generalise these methods for application to blended sources, described in chapter 5. This chapter discusses the methods used to obtain photometric redshifts in general. Section 4.1 details photometric redshift methods for determining the redshifts of individual sources; the two main methods for doing this, template-based and empirical methods, are discussed in sections 4.1.1 and 4.1.2 respectively. Section 4.2 then details methods for inferring the redshift distribution of a population of sources.

4.1 Photometric redshift Methods

As described above, photometric redshift methods are utilised for the analysis of cosmological galaxy surveys for two distinct uses, inferring the redshift of individual sources and inferring the redshift distribution of a population of sources. This section describes several photometric redshift methods for the first of these uses.

4.1.1 Template-based methods

A conceptually simple method to compute photometric redshifts is to frame the problem as a statistical regression by specifying the forward model of the data. In the case of photometric redshifts, the forward model for the observed fluxes is simply the redshifted spectrum of the source, integrated over the responses of each colour filter. These responses quantify the proportion of light that passes through each filter as a function of wavelength. While these responses are known for each filter, the same is not true of the intrinsic spectrum of the source. Instead, we assume that this intrinsic

spectrum is well-represented by one or more *templates*, model spectra that are specified *a priori*. Photometric redshift methods that determine redshifts in this way are therefore termed *template-based* methods.

Early approaches to template-based photometric redshifts (e.g., Loh and Spillar, 1986; Gwyn and Hartwick, 1996) were maximum-likelihood based, where a likelihood function is maximised w.r.t. the free parameters of the model. The free parameters for template-based photometric redshifts are the choice of template itself, the redshift, and a normalisation factor α. We label the model flux in a band b given by a template t redshifted to z as $T_{t,b}(z)$. The free parameters are then found by minimising

$$\chi^2 = \sum_b \frac{\left(F_b - \alpha T_{t,b}(z)\right)^2}{\sigma_b^2}\,, \tag{4.2}$$

where F_b and σ_b are the observed flux and the associated uncertainty in band b respectively. The template t is then chosen by computing χ^2 for all templates and choosing the one that gives the minimum χ^2. These maximum-likelihood methods are the basis for several widely used photometric redshift codes such as Hyperz (Bolzonella et al., 2000) and Le-PHARE (Ilbert et al., 2006).

Later template-based photometric redshift methods use Bayesian techniques, first introduced to the field by Benítez (2000), implemented in his Bayesian Photometric Redshifts (BPZ) code. Here, the output is not a maximum-likelihood estimate of the redshift but rather a posterior distribution of the redshift, given the observations. Bayesian methods are discussed more generally in section 3.1.

BPZ considers the flux data in the form of colours and the reference band magnitude. The colours are defined for band b to be $C_b \equiv {}^{F_b}/_{F_0}$, where F_0 is the flux observed in a designated *reference band*. These colours $\boldsymbol{C}$, alongside the magnitude in the reference band m_0, form the data on which the posterior distribution is conditioned. Modelling the colours rather than the flux itself allows us to avoid also constraining the normalisation of the template. This posterior is then given by marginalising over the discrete choice of template t, giving

$$\begin{aligned} P(z \mid \boldsymbol{C}, m_0) &= \sum_t P(z, t \mid \boldsymbol{C}, m_0) \\ &\propto \sum_t P(\boldsymbol{C} \mid z, t) P(z, t \mid m_0)\,. \end{aligned} \tag{4.3}$$

Developing the posterior in this way involves making the assumption that the likelihood depends only on the colours, and not on the reference band magnitude. This magnitude m_0 appears only in the prior, allowing the redshift and type distributions

to be magnitude dependent, and is assumed to have a negligible uncertainty as its true value is not marginalised over. Benítez (2000) specified a Gaussian likelihood for the colours. This is also an approximation, even if the fluxes have Gaussian uncertainties. Nevertheless, flux uncertainties are typically sufficiently small that these approximations are valid.

This choice of likelihood is also not integral to a Bayesian approach to template-based photometric redshifts. The large number of photons observed for a detected source suggest predominantly Gaussian statistics. However, non-Gaussian flux errors can arise from difficulties in performing photometry on crowded fields, an effect that can be successfully accounted for in photometric redshift applications by modifying the likelihood (e.g., Wittman et al., 2007).

Bayesian methods have several advantages over maximum likelihood approaches. Firstly, as described above, it is important the uncertainties arising from photometric redshifts are rigorously accounted for an are able to be propagated throughout the analysis. The posterior distributions resulting from Bayesian methods provide a way to quantify these uncertainties more generally than simply specifying a variance. Secondly, marginalising over templates as in equation 4.3 provides a way of incorporating the uncertainty over this choice into the final posterior.

Finally, the prior distributions present in Bayesian methods provide a mechanism by which astrophysical knowledge about the properties of galaxies can be included. For example, it can be specified that brighter galaxies are more likely to be low redshift than high redshift, and the very high redshift galaxies are rare in general. Including this prior has been found to reduce the number of catastrophic outliers compared to a maximum-likelihood approach (Benítez, 2000). Placing priors over other observable aspects of the source have also been considered. For example, using a surface brightness prior was found to reduce the number of outliers, the bias and the scatter of photometric redshifts applied to both ground-based and space-based observations of sources at redshifts $0.4 \leq z \leq 1.3$ (Stabenau et al., 2008).

The use of priors within a Bayesian inference is discussed more generally in section 3.1.7. In general, priors can be specified in a variety of ways, such as to express ignorance about a quantity as in the case of non-informative priors, or for mathematical convenience as in the case of conjugate priors. It should be noted, however, that the priors that are present in these Bayesian photometric redshift methods often directly correspond to physically meaningful quantities, e.g., the redshift distribution of the population.

The accuracy of template-based photometric redshift methods depends on the

templates. If the intrinsic spectra of the sources of interest are poorly-represented by the template set, the inferred redshift will inevitably be of low accuracy due to model misspecification. It is therefore important that the templates in the template set represent the types of sources being observed and are numerous enough to densely cover the range of possible galaxy types.

Template sets can be derived either from observations of real sources such as the commonly used CCW (Coleman et al., 1980) and Kinney et al. (1996) templates, or predicted by star-formation models (Brammer et al., 2008). It is also common practice to interpolate between templates in order to increase the coverage of the template set (Sánchez et al., 2014).

Methods have also been developed to address the potential mismatch between galaxy spectra and templates. The Zurich Extragalactic Bayesian Redshift Analyzer (ZEBRA, Feldmann et al., 2006) uses a set of training galaxies with known redshifts in in addition to their photometry in order to modify templates to be more representative using a regularised χ^2 minimisation method. The photometric redshift code EAZY (Brammer et al., 2008) also includes a template error function that accounts for template mismatch by iteratively fitting the redshifts and determining the flux residuals.

The work presented in chapter 6 is a generalisation of template-based photometric redshift methods to the case of blended sources.

4.1.2 Empirical Methods

Unlike template-based photometric redshift methods where the relationship between fluxes and redshift is specified *a priori* through a template set, empirical methods determine this relationship from data alone. Given a training set of sources with both photometry and known spectroscopic redshifts, these methods fit very flexible functions that map between the fluxes and redshifts.

Empirical photometric redshift methods have a history almost as long as template-based methods. The first empirical photometric redshifts were performed by Connolly et al. (1995), who used linear regression to predict redshifts from four ground-based optical bands using 254 spectroscopically observed galaxies. The predicted redshift was then given by the typical linear regression formula

$$z = \boldsymbol{\beta}^T \boldsymbol{X} \,, \tag{4.4}$$

where $\boldsymbol{\beta}$ is a vector of coefficients to be determined, and $\boldsymbol{X}$ is a vector of features,

i.e., functions of the data. Connolly et al. (1995) first tested linear features, where $X_{b+1} = m_b$, i.e., elements of the feature vector are simply the observed magnitudes. The first element of the feature vector is given by a constant term $X_1 = 1$, so that the corresponding coefficient β_1 acts as an intercept. This approach resulted in an RMS error of $\sigma_z = 0.057$. Connolly et al. (1995) also tested quadratic features where, in addition to the previous linear features, elements of $\boldsymbol{X}$ are products of two magnitudes $m_i m_j$. In this case, the RMS error reduced to $\sigma_z = 0.047$, a reduction of $\approx 20\%$.

While empirical methods have a long history, their usage has grown significantly with the development and introduction of machine learning techniques. The distinction between machine learning and other empirical methods is somewhat blurred, though modern machine learning methods such as neural networks (e.g., LeCun et al., 2012) tend to have many more free parameters than the linear regression model above, and so must be fitted with a considerably larger training dataset. However, the benefit to undergoing this extra effort is that the resulting model is more flexible, and thus more able to capture complex non-linear relationships between the fluxes and redshifts. Section 3.2 introduces machine learning methods more generally.

The first application of modern machine learning techniques to photometric redshift estimation was by Firth et al. (2003) who trained neural networks on 10^4 SDSS (Stoughton et al., 2002) galaxies at redshifts $z \lesssim 0.35$, achieving an RMS redshift error of $\sigma_z \approx 0.021$. Neural networks consist of a series of layers, each comprising many nodes. The values of the nodes in the first layer are set to the input data to the network. The nodes in subsequent layers are then set to the result of passing a linear combination of the previous layer's nodes through a non-linear function, known as the activation function $f(\dots)$. Thus, the value of node j in layer n is given by

$$x_{j,n} = f\left(\sum_i w_{i,m} x_{i,m}\right) , \tag{4.5}$$

where $x_{i,m}$ is node i in the previous layer m and $w_{i,m}$ is the corresponding weight. The nodes in the final layer of the network correspond to the output. The final layer in the case of photometric redshift estimation therefore generally consists of only a single node for the redshift[1].

Training a neural network involves finding the optimum values for the weights. This is done by specifying a *loss function* which quantifies the distance between the prediction of the network and the spectroscopic ground truth. The weights can then be

[1]While a single output network is most common, some neural network approaches to photometric redshifts output several values parametrising a PDF in redshift space (e.g., D'Isanto and Polsterer, 2018)

optimised iteratively using gradient-based optimisation methods; the process of recursively applying chain rule to obtain the derivative of each weight w.r.t. the loss function is known as *backpropagation*. Gradient-based optimisation methods are discussed in more detail in section 3.2.3.

The neural network-based photometric redshift codes ANNz (Collister and Lahav, 2004) and its successor ANNz2 (Sadeh et al., 2016) are now widely used in cosmlogical galaxy surveys such as the Dark Energy Survey (DES, Sánchez et al., 2014; Gschwend et al., 2018) and the Kilo-Degree Survey (KiDS, Bilicki et al., 2018). A variety of other machine learning methods have also been used for estimating photometric redshifts such as random forests (e.g., Carliles et al., 2010; Carrasco Kind and Brunner, 2013), boosted decision trees (e.g., Gerdes et al., 2010), support vector machines (e.g., Wadadekar, 2005) and Gaussian processes (e.g., Way and Srivastava, 2006; Almosallam et al., 2016).

An advantage of machine learning approaches is the ease with which they can be extended to include extra input features. These features can simply be included as extra inputs, provided that they are also available for the training set sources. For example, Collister and Lahav (2004) investigated using Petrosian radii as inputs to their model. The Petrosian radius (e.g., Blanton et al., 2001) is defined by measuring the average surface brightness of a galaxy within an annulus of a particular radius. The ratio of this value with the average surface brightness at smaller radii defines the Petrosian ratio, with the Petrosian radius being the radius at which this ratio equals a specified value. By including the 50% and 90% Petrosian radii of sources as additional inputs, Collister and Lahav (2004) found that the RMS error was reduced by $\approx 3\%$.

Soo et al. (2018) investigated the effect of including a variety of morphological parameters as input. They found that the fraction of outliers was reduced by 14% when utilising photometric data with five optical bands, and that the results for two photometric bands with morphological information were comparable to those with five photometric bands alone. Finally, machine learning methods designed for image analysis, such as convolutional neural networks (CNNs) can be used to construct photometric redshift methods that utilise entire galaxy images as input, rather than reducing this information to a vector of fluxes (e.g., D'Isanto and Polsterer, 2018).

Another benefit to machine learning methods is that any systematic effects in the photometry that is present in both the training and test sets is learnt as part of the training process. As a result, these effects are automatically accounted for in the final redshift estimates.

The data-driven approach of machine learning methods avoids the potential pit-

falls of template mismatch, but instead relies on the training set being representative. If this is the case, the accuracy of these methods can be greater than that of than template-based methods (Hildebrandt et al., 2010) In practice however, training sets are often shallower than the photometric sample. Redshift estimates of galaxies not represented by the training data are much less reliable (Beck et al., 2017). The common case where spectroscopic training data is shallower than the photometry can lead to biases where the redshifts of high redshift galaxies are underestimated (Rivera et al., 2018). In summary, machine learning methods are more effective at *interpolation* than at *extrapolation*.

As described above, an accurate understanding of uncertainties in an important aspect of photometric redshift methods, and these uncertainties are represented most generally by PDFs. Many machine learning methods are able to produce PDFs as their output. For example, ANNz2 (Sadeh et al., 2016) estimates the uncertainty due to photometry uncertainties by using K-nearest-neighbours to identify the most similar counterparts to the data in the training set and uses those to estimate redshift uncertainties. These are then averaged over an ensemble of models with different hyperparameters. Similar ensemble approaches can also be used between different decision trees in a random forest (Carrasco Kind and Brunner, 2013). Some machine learning methods, such as Gaussian process (e.g., Way and Srivastava, 2006; Almosallam et al., 2016) are intrinsically probabilistic and so provide PDFs without modification. Machine learning methods such as neural networks can also be trained to output parameters describing a PDF in the form of a Gaussian mixture model (e.g., D'Isanto and Polsterer, 2018), an approach known as *mixture density networks* (Bishop, 1994).

It is worth noting that despite drawing comparisons between template-based and empirical methods, these methods are not always so distinct in practice. For instance, the priors of Bayesian methods typically include a set of parameters that are fitted using a set of training data (e.g., Benítez, 2000; Schmidt and Thorman, 2013, also see section 6.1.6). In addition, recent applications of photometric redshifts have used hybrid methods that combine a template-based approach with machine learning methods (e.g., Speagle and Eisenstein, 2017; Leistedt et al., 2018; Duncan et al., 2018).

The work presented in chapter 7 is a machine-learning based photometric redshift method designed for inferring the redshifts of blended sources, but trained using a training set of unblended sources.

4.2 Inferring Redshift Distributions

The methods described above are designed for inferring the redshift of a single source. However, as detailed in section 2.3.3, another necessary input for doing cosmology with photometric galaxy surveys is the redshift distribution of the entire population of sources in each tomographic bin. We therefore need photometric redshift methods designed for this purpose.

A simple method for obtaining these distributions is *stacking*, where the redshift posterior distributions of each source are summed together (e.g., Bonnett et al., 2016), i.e.,

$$n(z) \equiv \sum_i P(z \mid \boldsymbol{F}_i) \tag{4.6}$$

where $\boldsymbol{F}_i$ is the flux vector of source i. However, estimating redshift distributions in this way doesn't provide a way to obtain uncertainties on the distributions, and can result in a poor recovery of the true distribution; see Leistedt et al. (2016) for a discussion of this problem.

Another type of photometric redshift method designed for inferring redshift distributions is *clustering redshifts*, introduced by Newman (2008). This method can be seen as somewhat distinct from both template-based and empirical photometric redshift methods as it does not utilise the colours of the source. Instead, the redshift distribution of the population is determined completely by its spatial distribution. By cross-correlating the positions of the target sources with another population of sources with known redshifts, the redshift distribution of the target sources can be be inferred statistically.

Various authors (e.g., Schmidt et al., 2013; Ménard et al., 2013; Rahman et al., 2016; Scottez et al., 2018; Bates et al., 2019) have further developed the method and have applied it on real data. McLeod et al. (2017) presented a method to jointly constrain clustering redshifts and cosmological parameters self-consistently. Computing clustering redshifts requires constraints or assumptions on the evolution of galaxy bias. To model this, Rau et al. (2019) developed a hierarchical Gaussian process model to jointly infer clustering redshifts and bias parameters.

Another recent method for inferring redshift distributions is known as *DIR calibration* (e.g., Hildebrandt et al., 2017) First developed by Lima et al. (2008), this is a reweighting method that modifies spectroscopic redshift distributions so that their photometric properties match those of the target population. This is done using a nearest-neighbours method that estimates the ratio of the density in magnitude-space of the spectroscopic objects with the corresponding magnitude space of the target popu-

lation. This ratio can then be used to reweight the spectroscopic objects. The resulting reweighted redshift distribution is then an estimate of the redshift distribution of the target population.

The DIR calibration method was used by the Kilo Degree Survey (KiDS) for inferring photometric redshifts in their cosmological weak lensing analysis (KiDS Hildebrandt et al., 2017). The template-based BPZ method (Benítez, 2000) was first used to separate sources into tomographic bins. The redshift distributions of these bins were then inferred using DIR. This is in contrast to the Dark Energy Survey (DES) which utilised the stacking method described above for their year-1 analysis (Hoyle et al., 2018), though clustering redshifts were also utilised for validation purposes. Recent work by DES (Buchs et al., 2019) describes a method based on self-organising maps to combine information from three distinct datasets; the target wide photometric dataset, a small dataset with well constrained redshifts typically obtained from spectroscopy, and a deep photometric sample. On simulated year-3 data, this method improved the error on the mean redshift of each tomographic bin by 60% over the year-1 analysis.

4.2.1 Bayesian hierarchical approach

A Bayesian hierarchical approach to inferring redshift distributions from photometric data was developed by Leistedt et al. (2016). Like the stacking method described above, this can be seen as a generalisation of existing template-based photometric redshift methods for single sources. However, in contrast to stacking, using a Bayesian hierarchical model is statistically rigorous and able to provide a full posterior distribution over redshift distributions for propagating uncertainties. This method works as follows.

While other Bayesian template-based methods such as BPZ (Benítez, 2000) use informative priors, Leistedt et al. (2016) take a non-parametric approach, specifying the joint redshift-magnitude-template prior as a three-dimensional histogram, i.e.,

$$P(t, z, m \mid \{f_{ijk}\}) = \sum_{ijk} \frac{f_{ijk}}{(z_{j,\max} - z_{j,\min})(m_{k,\max} - m_{k,\min})} \delta_{t,t_i} \times \Theta(z - z_{j,\min})\Theta(z_{j,\min} - z)\Theta(m - m_{k,\min})\Theta(m_{k,\min} - m)\,, \tag{4.7}$$

where δ is the Kronecker delta, $\Theta(\dots)$ is the Heaviside step function, and f_{ijk} is the coefficient that parametrises the height of bin ijk.

The goal of the inference is then to infer the joint distribution over the parameters $\{z_g, t_g, m_g\}$ for each galaxy g, and the histogram coefficients $\{f_{ijk}\}$, given the flux of

each galaxy $\{\boldsymbol{F}_g\}$. The joint posterior is therefore given by

$$P(\{z_g, t_g, m_g\}, \{f_{ijk}\} \mid \{\boldsymbol{F}_g\}) \propto P(\{f_{ijk}\}) \prod_g P(\boldsymbol{F}_g \mid z_g, t_g, m_g) P(z_g, t_g, m_g \mid \{f_{ijk}\}) \,, \tag{4.8}$$

where $P(\boldsymbol{F}_g \mid z_g, t_g, m_g)$ is the standard template-based likelihood, $P(z_g, t_g, m_g \mid \{f_{ijk}\})$ is given by equation 4.7, and the coefficient prior $P(\{f_{ijk}\})$ is given by a Dirichlet distribution; this choice of prior is discussed in more detail in section 8.1.3. Leistedt et al. (2016) describe an efficient Gibbs sampling method to draw samples from this joint distribution by describing sources in terms of the histogram bin they fall in, rather than their continuous properties directly. The likelihood distribution for the number of sources in each bin given their heights is then a multinomial distribution. The conjugate prior of this likelihood function is the Dirichlet distribution, allowing the conditional distributions to be sampled without rejection.

An extension to this approach was presented by Sánchez and Bernstein (2019). This uses a similar Bayesian hierarchical model to combine a template-based photometric redshift approach with the clustering redshifts discussed above. This is done by modelling the spatial distribution of galaxies as a doubly-stochastic Poisson process where galaxies are Poisson samples from an underlying density field that is itself stochastic. Gibbs sampling of the resulting posterior is made tractable by assuming redshift to be discrete and the stochastic density fields in each redshift slice to be independent. Alarcon et al. (2019) then apply an extension of this model to a set of N-body simulations. This is done by modelling the clustering probability using a kernel density estimator and allowing the galaxy bias[2] to be redshift dependent.

The work presented in chapter 8 is a similar Bayesian hierarchical approach to inferring redshift distributions of populations of blended sources. However, rather than using a template-based approach, our method uses a training set, generalising the Gaussian mixture model-based method for individual sources we present in chapter 7.

[2]The galaxy bias relates the underlying matter density to the density of the galaxies themselves.

Chapter 5

Blending

As discussed in section 2.3.5, future photometric galaxy surveys such as the Large Synoptic Survey Telescope (LSST, Ivezić et al., 2019) will, among various other science goals, constrain cosmological parameters to high precision. A factor contributing to this increased precision over previous surveys is the photometric depth of LSST, which will observe to a depth of $m_r < 27$ (LSST Science Collaboration et al., 2009).

The result of this increase in depth is an increase in the number density of sources observed compared to other wide area surveys, as sources that were previously too faint can be present in the sample. To illustrate this, we can look at n_{eff}, a weak lensing-specific measure of the effective number density of sources. This value was defined by Albrecht et al. (2006) to be the number density of sources with perfectly measured shear that would correspond to the same level of shear noise as the true sample of sources, the shapes of which are not measured perfectly.

The effective number density of sources n_{eff} was estimated by Chang et al. (2013) for LSST to be $n_{\text{eff}} \approx 31 - 36\,\text{arcmin}^{-2}$, compared to $n_{\text{eff}} \approx 17\,\text{arcmin}^{-2}$ for CFHTLens (Heymans et al., 2012), an existing weak lensing survey. This significant increase in the number density of sources, combined with an area coverage of $\approx 18000\,\text{deg}^2$, drives the increase in precision of cosmological parameter constraints expected for future surveys like LSST. However, since LSST is a ground-based survey, this also introduces a new problem that was negligible previously, known as *blending*.

Blending refers to observed sources overlapping to some degree when projected onto the sky. Examples of blended sources identified in the GAMA blended sources catalogue (Holwerda et al., 2015) are shown in Figure 5.1. The extent to which this overlap occurs impacts the effect it has on cosmological measurements. When two sources are overlapped only slightly with their centres otherwise well separated, blending can be easily identified. In contrast, this overlap can also be to such a degree that

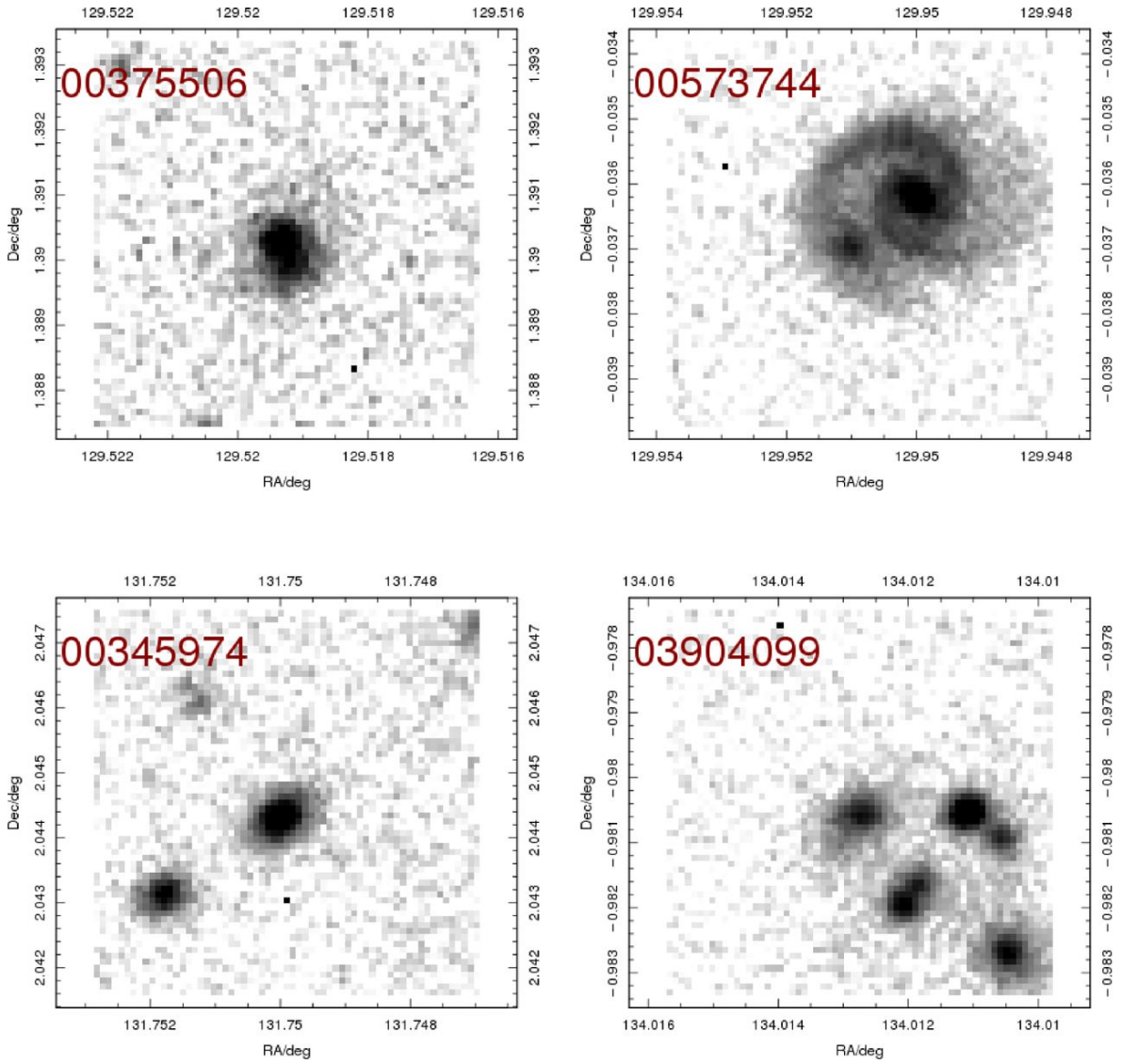

Figure 5.1: Four examples of blended sources identified by the GAMA blended sources catalogue (Holwerda et al., 2015). Images are taken in the SDSS *i*-band (Stoughton et al., 2002), and the red numbers are their corresponding GAMA survey (Baldry et al., 2017) ID-numbers. *Figure reproduced with permission from Holwerda et al. (2015).*

blending cannot be successfully identified and thus the source is detected as a single source only. This is known as *ambiguous blending* (Dawson et al., 2016). Throughout this thesis, we refer to the underlying physical galaxies comprising a blended source as *constituents*[1].

The increased number density of sources in future cosmological galaxy surveys such as LSST means that blending will impact a significant proportion of sources. Dawson and Schneider (2014) estimate that $45 - 55\%$ of sources will be blended to

[1]It is common in deblending literature to refer to the physical galaxies in a blended sources as *components*. However, chapters 7 and 8 make use of Gaussian mixture models, where the individual Gaussian distributions are also typically referred to as components. To avoid confusion, we thus decide to follow the latter usage of the term, and use *constituent* to refer to blended galaxies instead.

some degree in LSST, with $15 - 20\%$ of all sources being ambiguous blends. Blending will therefore have a significant effect on inferences of cosmological parameters. For example, incorrectly identifying multiple blended sources as a single source can have an impact on shape measurements. Dawson et al. (2016) estimate that ambiguous blends will result in a $\approx 14\%$ increase in shear noise. The superposition of fluxes from blended sources is also expected to increase the rate of catastrophic failures of photometric redshifts (Mandelbaum, 2018). Inferring photometric redshifts from blended sources is the subject of the research work of this thesis.

As a result of the large effect caused by blended sources, several methods to separate these sources into their individual constituents have been developed. These *deblending* methods are discussed in section 5.1. Section 5.2 then describes some difficulties with deblending that we aim to address in this thesis.

5.1 Deblending Methods

This section details a chronological development of deblending methods utilised by cosmological galaxy surveys.

5.1.1 Automatic source extraction

In order to efficiently make use of the large datasets obtained from cosmological galaxy surveys, automated techniques for identifying and extracting measurements from sources are necessary. COSMOS deblender (Beard et al., 1990) and Source Extractor (SExtractor, Bertin and Arnouts, 1996) are two similar methods for doing this. The latter is very commonly used throughout astronomy; we detail this method below.

Firstly, the background level of the sky is estimated. This is done through a process of *sigma clipping*, where pixels are iteratively removed until the values of all remaining pixels are within three standard deviations of the median. Then, if the final standard deviation is within 20% of its original value, the background level is taken to be the mean of the remaining pixels. If not, the background level is given by $2.5 \times \text{median} - 1.5 \times \text{mean}$. The values in this process were chosen as they were found to give accurate results on simulated test data (Bertin and Arnouts, 1996). This process is repeated over many regions of the image independently, and the results linearly interpolated between to construct a background map that estimates the level of background flux in each pixel.

SExtractor then uses a multiple-thresholding technique to identify and separate sources. The image is thresholded at several different levels, and contiguous regions of pixels above the threshold are noted at each level. As the threshold is increases, regions of pixels that were contiguous become separated by pixels that fall below the new threshold. If the flux contained in these separated regions is over a specified fraction of the total flux within the initially contiguous region, these regions are separated into distinct subregions. If this fraction is too small, the thresholding-and-splitting continues until this condition is met.

In this way, all pixels within the image are assigned to either the background or a subregion. These regions must then be assembled into separated sources. To do this, each subregion is fitted with a two-dimensional Gaussian profile. This profile is then used to statistically assign pixels to sources.

While this method is an efficient way to identify and extract sources from an image automatically, it is not well suited to deblending highly overlapping sources. The reason for this is that the entire flux of each pixel is assigned to a single source only. This is a poor model when sources are highly-overlapping, as the flux from each pixel is a combination of all sources in that direction on the sky.

5.1.2 Fractional splitting of pixel fluxes

As described above, a deblending method that separates sources by assigning the entire flux of each pixel to a single source is not a good model for highly-overlapping fields. To combat this, several methods (e.g., Weir et al., 1995; Fukugita et al., 1995) have been developed which split the flux of each pixel between several sources. One of these methods is the SDSS deblender (Lupton, 2005), a method used to deblend images from the Sloan Digital Sky Survey (SDSS, Stoughton et al., 2002).

The SDSS deblender works by constructing galaxy profile templates from images by assuming a symmetry ansatz. A list of peak pixels is first identified in the image, i.e., maxima pixels surrounded by pixels of a lower flux. This list is then reduced based on a series of criteria, such as a specified minimum distance between peaks. Peaks of flux F_p should also not be connected by a contiguous region of pixels with fluxes greater than $F_p - 3B$, where B is the background flux.

Given a peak r corresponding to a single constituent galaxy, a profile template $T_{r,i}$ is then constructed over pixels indexed by i. This is done by considering pairs of pixels i and j with fluxes F_i and F_j positioned symmetrically relative to the peak, and

setting

$$T_{r,i} = T_{r,j} = \min(F_i, F_j)\,. \tag{5.1}$$

Doing this for all pixels allows the construction of a symmetric template $T_{r,i}$ without *a priori* specifying a particular profile.

The model for the flux F_i of pixel i is then given by the weighted sum over templates, i.e.,

$$F_i \approx \sum_r w_r T_{r,i}\,, \tag{5.2}$$

where the weights are found by minimising the sum of the squared errors over pixels E, given by

$$E = \sum_i \left(F_i - \sum_r w_r T_{r,i} \right)\,. \tag{5.3}$$

The total flux is then fractionally split between all constituents r according to the weights w_r, with the total flux being conserved. Thus, the flux corresponding to constituent r in pixel i is given by

$$F_{r,i} = F_i \frac{w_r T_{r,i}}{\sum_{r'} w_{r'} T_{r',i}}\,. \tag{5.4}$$

In order to apply this deblending method to multi-band data, the peak pixels are constrained to be in the same position over all bands. However, the templates in each band are constructed independently of the other bands, meaning that the method is fundamentally monochromatic. Deblending methods that utilise the colour information across multiple bands are discussed in the next section.

5.1.3 Going beyond monochromatic deblending

Cosmological galaxy surveys observe sources in a variety of wavelength bands. The resulting poly-chromatic data can potentially be very informative for the purpose of deblending, as two galaxies with very different spectra can be identified from their colours, even if they are closely overlapping. One deblending method which takes account of this information is MuSCADeT (Joseph et al., 2016). This method constructs non-parametric models of galaxy profiles in a wavelet basis, enforcing that the resulting solution is sparse in this basis to reduce degeneracies and ensure convergence.

An extension to the MuSCADeT method designed for LSST (Ivezić et al., 2019) observations is `scarlet` (Melchior et al., 2018). This method also allows non-parametric fits to the colours of each constituent, rather than specifying them *a priori* as in MuSCADeT. This method works as follows.

We label the matrix of pixel fluxes of a multi-band survey as $\underline{\boldsymbol{M}} \in \mathbb{R}^{B \times N}$, where B is the number of filter bands in the survey, each image consists of N pixels. This is modelled as a matrix product over k constituents, given be

$$\underline{\boldsymbol{M}} = \underline{\boldsymbol{A}}\underline{\boldsymbol{S}}\,, \tag{5.5}$$

where each column in $\underline{\boldsymbol{A}} \in \mathbb{R}^{B \times k}$ is the photometric amplitude of a constituent k, i.e., the flux that would be observed if this constituent were not blended, and each row in $\underline{\boldsymbol{S}} \in \mathbb{R}^{k \times N}$ represents the spatial distribution of a constituent k. Both matrices $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{S}}$ are constrained to have all elements positive; equation 5.5 thus represents a *non-negative matrix factorisation*, a common dimensionality-reduction technique (e.g., Hoyer, 2004).

Like the SDSS deblender described in section 5.1.2, additional heuristic constraints are placed on the values of $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{S}}$. Firstly, a symmetry constraint is placed on pairs of pixels i and j surrounding the peak of the profile so that $S_{k,i} = S_{k,j}$. Secondly, the spatial distribution is constrained so that the resulting profile monotonically decreases in flux radially outwards from the peak. Finally, penalties are imposed on the norms of $\underline{\boldsymbol{A}}$ and $\underline{\boldsymbol{S}}$ to encourage sparse solutions.

The matrix factorisation is computed by minimising, subject to the constraints above, the square of the Frobenius norm between the model $\underline{\boldsymbol{M}}$ and the observed pixels $\underline{\boldsymbol{Y}}$, i.e.,

$$f(\underline{\boldsymbol{A}}, \underline{\boldsymbol{S}}) = \frac{1}{2}\left(\|\underline{\boldsymbol{Y}} - \underline{\boldsymbol{A}}\underline{\boldsymbol{S}}\|_2\right)^2 \tag{5.6}$$

where the Frobenius norm is defined as

$$\|\underline{\boldsymbol{X}}\|_2 = \sqrt{\sum_i \sum_j |X_{i,j}|^2}\,. \tag{5.7}$$

However, these constraints are non-differentiable, presenting a problem for typical constrained optimisation methods. To counter this, `scarlet` utilises *proximal operators*, a method for performing constrained optimisation of a function $f(\boldsymbol{x})$ with non-differentiable constraints. If these constraints $g(\boldsymbol{x})$ are defined such that the optimisation target is $f(\boldsymbol{x}) + g(\boldsymbol{x})$, the proximal operator, parametrised with the step-size λ, is given by

$$\mathrm{prox}_{\lambda_g}(\boldsymbol{x}) \equiv \underset{\boldsymbol{u}}{\mathrm{argmin}} \left\{ g(\boldsymbol{u}) + \frac{1}{2\lambda}\left(\|\boldsymbol{x} - \boldsymbol{u}\|_2\right)^2 \right\}. \tag{5.8}$$

Since the proximal operators of many types of non-differentiable constraints can be evaluated analytically, formulating in the problem in this way side-steps the issue of non-differentiability. The value of $\boldsymbol{x}$ can then be updated iteratively using a gradient

descent step, given at iteration $i + 1$ by

$$\boldsymbol{x}^{i+1} = \text{prox}_{\lambda_g}(\boldsymbol{x}^i - \lambda \nabla f(\boldsymbol{x}^i)) \,. \tag{5.9}$$

The result of this optimisation procedure converges to the value of $\boldsymbol{x}$ that minimises the function subject to the constraints, as desired.

Alternative methods for poly-chromatic deblending include machine learning-based methods such as convolutional neural networks (e.g., Reiman and Göhre, 2019; Burke et al., 2019) which are designed for image processing tasks. These methods learn to deblend images by using a training set of blended sources and the corresponding constituent images. These training sets can be constructed from a set of unblended galaxy images by artificially combining them. Machine learning methods are discussed in more detail in section 3.2.

5.2 Difficulties with Deblending

The deblending methods described above are commonly used in cosmological galaxy surveys to analyse blended sources. One advantage of these methods is the ease with which they can be incorporated into existing data analysis pipelines. Once a source has been deblended into several images, each of these constituents can be analysed in the same way as any other galaxy image.

However, analysing deblended sources in this way presents a problem for propagating uncertainties. Once the constituent images are separated, analysing them independently inevitably means that correlations between their fluxes are neglected. Put another way, accounting for these correlations necessitates utilising analysis methods specifically designed for blended sources, even if these sources have been deblended first. Neglecting these correlations could be problematic since, as the total flux of the blended source is well constrained by observations, we expect a potentially large correlation between the fluxes of each constituent. This is because the deblended flux in one constituent can be traded for another constituent while remaining consistent with the data. In addition, obtaining accurate uncertainties from these deblending methods can be difficult (see, e.g., Melchior et al., 2018).

The photometric redshift methods described in the research work of this thesis use an alternative approach. Rather than deblending sources and analysing each constituent separately, the analysis is performed on the blended data directly. In the case of photometric redshifts, these data are fluxes obtained from aperture photometry,

though in principle this same approach could utilise images of source themselves by constructing a forward model of these images. This is discussed in more detail in chapter 9. While this approach requires modifications to existing photometric redshift algorithms, it provides a rigorous statistical method for accounting for all uncertainties, including correlations between the constituents. Accounting for these uncertainties properly is important for ensuring that the final posterior distributions over cosmological parameters are an accurate representation of our state of knowledge.

Part II

Research

Chapter 6

Bayesian Photometric Redshifts of Blended Sources

This chapter is heavily based on work from Jones and Heavens (2019a).

The deblending methods described in chapter 5 that produce a set of component-separated maps are useful for later applying existing photometric redshift methods designed for individual components to. However, as described in section 5.2, splitting the analysis in this way can lose uncertainty information, such as the correlation between deblending parameters and the parameters in a subsequent analysis. A photometric redshift method that jointly constrains parameters directly from blended data provides a self-consistent, principled way to characterise and propagate this information.

In this chapter, we present a method that generalises the Benítez (2000) Bayesian photometric redshift (BPZ) method to the case of blended observations. This is a template-based method where the task of determining the component redshifts is cast as a Bayesian parameter inference problem. The product of such an inference is a joint posterior distribution of the redshift and magnitude of each component in the blended source. This distribution characterises the complete statistical uncertainty in the result in a way that can be propagated through the rest of the cosmological analysis. Determining the number of components in an observed source, i.e., whether or not it is blended, is treated as a model comparison problem. In this way, our method allows the identification of blended sources from aperture photometry alone.

A summary of our notation throughout this chapter is provided in Table 7.1. For parameters defined for each constituent in a source, we index over constituent using greek letters and indicate the collection of these using sets, i.e., $\{\theta\} \equiv \{\theta_\alpha, \theta_\beta, \dots \theta_N\}$. Vector quantities defined for each filter band are in bold $\boldsymbol{q}$, and observed quantities

are denoted with a hat $\hat{q}$. Where necessary, quantities defined for a specific number of constituents are distinguished by a subscript number in brackets, i.e., $q^{(1)}$ is the definition of q for a single constituent.

This chapter is organised as follows. In section 6.1, we describe our formalism for estimating redshifts as a parameter inference problem, describing its application to partially blended systems in section 6.2. In section 6.3, we discuss our inference methods, detailing how we use model comparison to identify blended objects in section 6.3.1. In section 6.4, we test our method on simulated observations. Section 6.5 describes a test of our method on the Galaxy And Mass Assembly survey (GAMA, Baldry et al., 2017) blended sources catalogue (Holwerda et al., 2015), for which spectroscopic redshifts are available. We conclude in section 6.6.

6.1 Blended photo-z formalism

6.1.1 Flux model

In the same way as other template-based photometric redshift methods, we assume that each observed constituent is well represented by one of a set of T templates. Each template t is defined by its rest-frame spectral flux density $F_t(\lambda_{\rm em})$ as a function of the emitted wavelength $\lambda_{\rm em}$. This template is redshifted and observed through a broadband filter b, the response of which is denoted $W_b(\lambda_{\rm obs})$ as a function of observed wavelength $\lambda_{\rm obs}$.

The flux of template t, at redshift z and observed in band b is then given by

$$T_{t,b}(z) = \frac{1}{cg^{\rm AB}C_b}\int_0^\infty F_t\left(\frac{\lambda}{1+z}\right)W_b(\lambda)\lambda {\rm d}\lambda\,, \tag{6.1}$$

where $g^{\rm AB} = 3631$ Jy is the zero-point of the AB-magnitude system and the normalisation $C_b \equiv \int_0^\infty \frac{W_b(\lambda)}{\lambda}{\rm d}\lambda$. By including $g^{\rm AB}$, our fluxes are dimensionless throughout, and the conversion between magnitudes and fluxes defined in the way is given by $F \equiv 10^{-0.4m}$. This template is then scaled by a normalisation a so that the flux of an object modelled with template t, at a redshift z and observed in band b is given by

$$F^{(1)}_{t,b}\big(z,a\big) = aT_{t,b}\big(z\big)\,. \tag{6.2}$$

We model the flux of blended sources as a linear combination of individual constituent

Table 6.1: A summary of the notation used throughout this chapter.

Symbol	Description
N	Number of constituents
T	Number of templates
B	Number of filter bands
z_α	Redshift of constituent α
$m_{0,\alpha}$	Reference band magnitude of constituent α
t_α	Template index of constituent α
$\{z\}$	Set of redshifts of each constituent
$\{m_0\}$	Set of reference band magnitudes of each constituent
$\{t\}$	Set of template indices of each constituent
b	Index over filter bands
b_0	Index of reference band filter
$\hat{F}_0$	Observed flux in reference band
$\hat{\boldsymbol{F}}$	Vector of observed fluxes, excluding the reference band
σ_0	Error on the reference band flux
σ_b	Error on the flux in band b
$F^{(1)}_{t,b}\big(z, m_0\big)$	Model flux for a single constituent in band b, at redshift z, with reference band magnitude m_0 and templates t
$F^{(N)}_{\{t\},b}\big(\{z\},\{m_0\}\big)$	Model flux for N-constituent blended source in band b, at redshifts $\{z\}$, with reference band magnitudes $\{m_0\}$ and templates $\{t\}$
χ	Set of cosmological parameters Ω_m, Ω_Λ and H_0
$\xi^{(N)}_\chi\big(\{z\}\big)$	Combination of up to N-point correlation functions describing the extra probability of N galaxies jointly sitting at redshifts $\{z\}$ due to clustering

fluxes. For a blend of N constituents, the flux observed in band b is given by

$$F^{(N)}_{\{t\},b}\big(\{z\},\{a\}\big) = \sum_{\alpha=1}^{N} a_\alpha T_{t_\alpha,b}\big(z_\alpha\big)\,, \tag{6.3}$$

where a_α is the normalisation for constituent α. For the reasons specified in section 6.1.3, we sample $m_{0,\alpha}$, the apparent magnitude of each constituent in the reference band b_0 rather than this normalisation directly. The normalisation a_α is then defined such that the model flux in the reference band is equal to $m_{0,\alpha}$. Thus, the model flux is given by

$$F^{(N)}_{\{t\},b}\big(\{z\},\{m_0\}\big) = \sum_{\alpha=1}^{N} \frac{10^{-0.4m_{0,\alpha}}}{T_{t_\alpha,b_0}\big(z_\alpha\big)} T_{t_\alpha,b}\big(z_\alpha\big)\,. \tag{6.4}$$

6.1.2 Fully-blended posterior

For a fixed number of constituents, photometric redshift determination is a parameter inference problem; we wish to infer the joint posterior distribution of the redshifts and apparent magnitudes of each constituent given a data vector $\hat{\boldsymbol{D}}$ of B broadband fluxes. This data vector is split into two parts $\hat{\boldsymbol{D}} = (\hat{\boldsymbol{F}}\,, \hat{F}_0)$, where $\hat{F}_0$ is the flux of the reference band and $\hat{\boldsymbol{F}}$ is the vector of the remaining $B-1$ fluxes. This is done since the normalisation of each constituent is defined in the reference band, and it is the flux of this band on which the priors are conditioned.

Following BPZ (Benítez, 2000), we set the flux of non-detections to zero. Likewise, bands that are not observed are given a flux of zero, with the corresponding error set to an extremely large value. As discussed in section 6.1.4, we assume that sources are selected using a magnitude limit on a single selection band. We therefore require that the source is detected in this band, by definition.

We start by writing our desired posterior as a marginalisation over templates for each constituent. For N constituents, we marginalise over sets of N template indices $\{t\}_i = \{t_\alpha, t_\beta \ldots t_N\}_i$. Each template index can take a value $1 \leq t \leq T$ and constituents may share the same template, so there are T^N of these sets to marginalise over, giving

$$P\Big(\{z\},\{m_0\}\,\Big|\,\hat{\boldsymbol{F}},\hat{F}_0,\chi,N\Big) = \sum_{i=1}^{T^N} P\Big(\{z\},\{t\}_i,\{m_0\}\,\Big|\,\hat{\boldsymbol{F}},\hat{F}_0,\chi,N\Big)\,. \tag{6.5}$$

We have emphasised that our posterior is defined for a fixed number of constituents by conditioning on N. In the general case where this number is unknown *a priori*, it can be inferred from the data; this is discussed in section 6.3.1. We have also

made the dependence on cosmological parameters, which are required for converting between distance and redshift, explicit in the above expression. These parameters are denoted by $\chi = \{\Omega_\mathrm{m}, \Omega_\Lambda, H_0\}$ for brevity. Applying Bayes rule, the posterior becomes

$$P\Big(\{z\}, \{m_0\} \,\Big|\, \hat{\boldsymbol{F}}, \hat{F}_0, \chi, N\Big) \propto \sum_{i=1}^{T^N} P\Big(\hat{\boldsymbol{F}}, \hat{F}_0 \,\Big|\, \{z\}, \{t\}_i, \{m_0\}, N\Big) P\Big(\{z\}, \{t\}_i, \{m_0\} \,\Big|\, \chi, N\Big) \,. \tag{6.6}$$

Since only the prior is dependent on cosmological parameters, we have removed the conditioning on χ from the likelihood. We then factorise the likelihood so that it is split in the same way as the data vector, giving

$$\begin{aligned} P\Big(\{z\}, \{m_0\} \,\Big|\, \hat{\boldsymbol{F}}, \hat{F}_0, \chi, N\Big) \propto \sum_{i=1}^{T^N} P\Big(\hat{\boldsymbol{F}} \,\Big|\, \{z\}, \{t\}_i, \{m_0\}, N\Big) \times \\ P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big) P\Big(\{z\}, \{t\}_i, \{m_0\} \,\Big|\, \chi, N\Big) \,. \end{aligned} \tag{6.7}$$

Since the magnitude of each constituent in the reference band is a sampled parameter in the posterior, our model for the reference band flux is simply the sum of these after converting from magnitudes to fluxes. As a result, the conditioning on $\{z\}$ and $\{t\}_i$ in the reference band likelihood is unnecessary and so has been removed. We assume that the error on the observed reference band flux is normally distributed with variance σ_0^2. Thus, the reference band likelihood is given by

$$P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left[\frac{-\left(\hat{F}_0 - \sum_{\alpha=1}^{N} 10^{-0.4 m_{0,\alpha}}\right)^2}{2\sigma_0^2}\right] , \tag{6.8}$$

where $m_{0,\alpha}$ is the sampled reference band magnitude for constituent α. Similarly, we use an uncorrelated multivariate Gaussian likelihood for $\hat{\boldsymbol{F}}$,

$$P\Big(\hat{\boldsymbol{F}} \,\Big|\, \{z\}, \{t\}_i, \{m_0\}, N\Big) = \prod_{b=1}^{B} \frac{1}{\sqrt{2\pi\sigma_b^2}} \exp\left[-\frac{\left(\hat{F}_b - F^{(N)}_{\{t\}_i,b}\big(\{z\}, \{m_0\}\big)\right)^2}{2\sigma_b^2}\right] , \tag{6.9}$$

where $F^{(N)}_{\{t\},b}\big(\{z\}, \{m_0\}\big)$ is the model flux is specified in equation 6.4 and σ_b^2 is the variance on the observed flux in band b.

The use of Gaussian likelihoods is shared with Benítez (2000). Measuring the flux of an object is effectively an exercise in counting photons; we might therefore expect the flux to be Poisson distributed. However, in the limit of a large number of counts, the Poisson distribution can be well approximated by a Normal distribution, justifying

this choice of likelihood. We also note that this specific choice of likelihood is not central to the method presented in this chapter. These distributions could be replaced by a different choice without impacting the rest of the formalism, though the results and implementation presented throughout assume Gaussian likelihoods as above.

6.1.3 Separating the joint prior

We now develop the prior so that it can be written in terms of individual constituents. We start by separating the joint prior into a product over priors on redshift, template and magnitude. Removing unnecessary conditioning, the joint prior becomes

$$P\Big(\{z\},\{t\},\{m_0\}\,\Big|\,\chi,N\Big) = P\Big(\{z\}\,\Big|\,\{t\},\{m_0\},\chi,N\Big)P\Big(\{t\}\,\Big|\,\{m_0\},N\Big)P\Big(\{m_0\}\,\Big|\,N\Big)\,. \tag{6.10}$$

This splitting up of the joint prior is similar to the approach of Benítez (2000). There are two important differences, however. Firstly, we include a prior on the apparent magnitude of each constituent. This differs from the approach of Benítez (2000) who considers the magnitude on which the redshift and template priors are conditioned to be exactly the observed reference band magnitude. The uncertainty in the scaling of the template is then represented by marginalising over a normalisation factor with an assumed flat prior. However, while this normalisation is not defined as such, it is acting to set the apparent magnitude of the source in the reference band. This magnitude is a quantity about which prior information is known.

The prior information on the apparent magnitude of constituents is particularly important in the blended case, as we need to consider more than just the overall magnitude of the source. The individual magnitudes of each constituent are necessary for scaling the model fluxes when predicting the model flux $F^{(N)}_{\{t\},b}\big(\{z\},\{m_0\}\big)$. In addition, motivated by existing galaxy observations and following Benítez (2000), our redshift and template priors for each constituent are magnitude-dependent. The individual constituent magnitudes are not directly observed in the blended case, and must therefore be considered as random variables in our model.

An alternative to sampling the magnitudes directly would be to make the fraction each constituent contributes the total flux a model parameter. However, the combination of intrinsic magnitude distributions and survey-specific selection effects would give the distribution of this fraction a highly complicated shape. Instead, including a prior on the magnitude of each constituent allows these effects to be easily accounted for.

The other important difference in the blended case is that each term in equation 6.10 is a joint prior over all constituents in the source. The redshift, type and magnitude properties of individual galaxies are much more well studied than those of blended sources. To make use of this information, we write these joint priors in terms of priors on the individual constituents.

Firstly, we assume that the template priors for each constituent are independent, i.e., galaxy types are not correlated. This allows us to split the template prior as

$$P\Big(\{t\}\,\Big|\,\{m_0\},N\Big)=\prod_{\alpha=1}^{N}P\Big(t_\alpha\,\Big|\,m_{0,\alpha}\Big)\,. \tag{6.11}$$

We also make the assumption that the redshift of each constituent depends only on its own type, not the types of other constituents. The redshifts of each constituent cannot be assumed to be independent however, as galaxies are distributed in a correlated way. The additional probability of finding N galaxies within a separation r over a random Poisson process is described by galaxy correlation functions of up to order N (Peebles, 2001). We denote the combination of correlation functions describing this extra correlation as $\xi_\chi^{(N)}\big(\{z\}\big)$, i.e., the excess probability for two galaxies is given by

$$1+\xi_\chi^{(2)}(z_\alpha,z_\beta)\equiv 1+\xi(r_{\alpha\beta})\,, \tag{6.12}$$

where the separation $r_{\alpha\beta}\equiv|\vec{r}_\alpha-\vec{r}_\beta|$ is the comoving distance between constituents α and β. In the two-constituent case, only the two point correlation function $\xi(r)$ is necessary. However for three or more galaxies, higher order correlation functions are needed, i.e.,

$$1+\xi_\chi^{(3)}(z_\alpha,z_\beta,z_\gamma)\equiv 1+\xi(r_{\alpha\beta})+\xi(r_{\beta\gamma})+\xi(r_{\alpha\gamma})+\zeta(r_{\alpha\beta},r_{\beta\gamma},r_{\alpha\gamma})\,, \tag{6.13}$$

where $\zeta(r_{\alpha\beta},r_{\beta\gamma},r_{\alpha\gamma})$ is the connected three-point galaxy correlation function.

The excess probability term $\xi_\chi^{(N)}\big(\{z\}\big)$ is defined in the posterior as a function of the constituent redshifts $\{z\}$, though the galaxy correlation function ξ (and higher order correlations) are defined in terms of comoving separation r. We therefore need to convert between the redshifts of each constituent and the comoving distance separating them. The line of sight comoving distance as a function of redshift is given by (e.g., Hogg, 1999)

$$r(z)=\frac{c}{H_0}\int_0^z\frac{\mathrm{d}z'}{E(z')}\,, \tag{6.14}$$

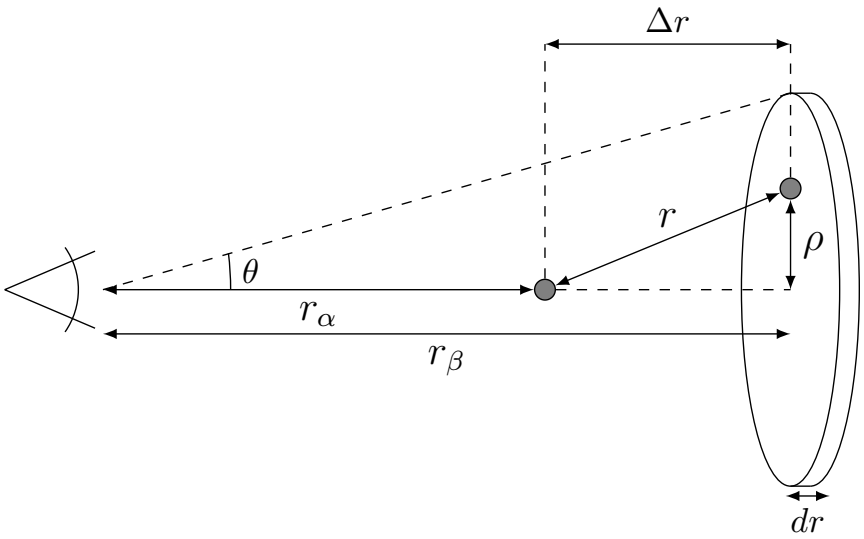

Figure 6.1: Diagram showing the setup of the $\xi_{\rm eff}$ calculation. We assume that two galaxies, represented by grey circles, will be blended if their angular separation is within θ. Given that these two galaxies are blended, the galaxy at a comoving distance r_β will lie within the disc.

where, neglecting radiation density and rewriting $\Omega \equiv \Omega_{\rm m} + \Omega_\Lambda$,

$$E(z) = \sqrt{\Omega_{\rm m}(1+z)^3 + (1-\Omega)(1+z)^2 + \Omega_\Lambda}\,. \tag{6.15}$$

We assume a flat Planck[1] (Planck Collaboration et al., 2016) cosmology throughout; $\Omega_{\rm m} = 0.3065$, $\Omega_\Lambda = 0.6935$ and $H_0 = 67.9\,{\rm km\,s^{-1}\,Mpc^{-1}}$.

However, the comoving distance separating constituents will depend not only on their redshifts, but also on their angular separation on the sky. As a result, we derive an effective correlation function $\xi_{\rm eff}$ that takes this angular dependence into account.

Consider the case of a two-constituent blend, as shown in Figure 6.1. The two constituents are at comoving distances r_α and r_β from the observer, with separation $\Delta r \equiv r_\beta - r_\alpha$. From the definition of the correlation function, we can write the ratio of the expected number of galaxies in a region with clustering N^ξ and that without N^0 as

$$1 + \xi_{\rm eff} = \frac{N^\xi}{N^0}\,. \tag{6.16}$$

Given that these constituents are blended, there is some maximum angular separation θ between them; we assume this to be small. We therefore compare the expected number of galaxies in a disc of width dr and radius $\rho_{\rm max} = r_\beta\theta$. The expected number

[1]We use the TT + lowP + lensing + ext values from Table 4.

without clustering is given by

$$N^0_{\rm disc} = \bar{n}\pi r_\beta^2 \theta^2 {\rm d}r \,. \tag{6.17}$$

To find the expected number with clustering, we integrate over the disc using the volume element of an annulus with radius ρ, i.e.,

$$N^\xi_{\rm disc} = \int_{\rho=0}^{r_\beta\theta} \bar{n}[1+\xi(r)]\, 2\pi\rho\, d\rho\, dr \,. \tag{6.18}$$

Thus, writing $r = \sqrt{\Delta r^2 + \rho^2}$, the ratio becomes

$$1 + \xi_{\rm eff} = \frac{2}{r_\beta^2\theta^2} \int_{\rho=0}^{r_\beta\theta} \left[1 + \xi\left(\sqrt{\Delta r^2 + \rho^2}\right)\right]\, \rho\, d\rho \,. \tag{6.19}$$

As described below, the effect of clustering is small. As a result, we adopt a simple power law for the two point correlation function,

$$\xi(r) \propto \left(\frac{r}{r_0}\right)^{-\gamma} \,. \tag{6.20}$$

Inserting this into equation 6.19 and integrating, the effective correlation function is given by

$$\xi_{\rm eff}(r_\alpha, r_\beta) = \frac{r_0^2}{\left(1-\frac{\gamma}{2}\right) r_\beta^2\theta^2} \left[\left(\frac{\Delta r^2 + r_\beta^2\theta^2}{r_0^2}\right)^{1-\frac{\gamma}{2}} - \left(\frac{\Delta r^2}{r_0^2}\right)^{1-\frac{\gamma}{2}}\right] \,. \tag{6.21}$$

The effect of the strength of clustering evolving with redshift can be included in this formalism by allowing the parameters r_0 and γ to vary with redshift (e.g., Sołtan, 2016). We test the effect of this on the redshift inference by using a toy model where $\gamma = 1.92$ is kept constant, while r_0 linearly varies between $r_0 = 5\,{\rm Mpc}\,h^{-1}$ at redshift $z = 2$ and $r_0 = 6\,{\rm Mpc}\,h^{-1}$ at redshift $z = 0.5$, with a linear extrapolation outside of this range. Since the value of $\xi_{\rm eff}$ is non-negligible only when $z_\alpha \approx z_\beta$, this interpolation of r_0 is evaluated using z_α only.

We then simulated two-constituent blends from a prior with $\xi_{\rm eff}$ included as described in section 6.4. Results assuming $\xi_{\rm eff} = 0$ showed negligible differences from those where the effect was included. At the population level, the RMS scatter defined in equation 6.45 changed by 0.205% between results including and excluding the correlation function. There were also negligible changes to the results at the individual source level. A comparison of maximum *a posteriori* results in each case are shown

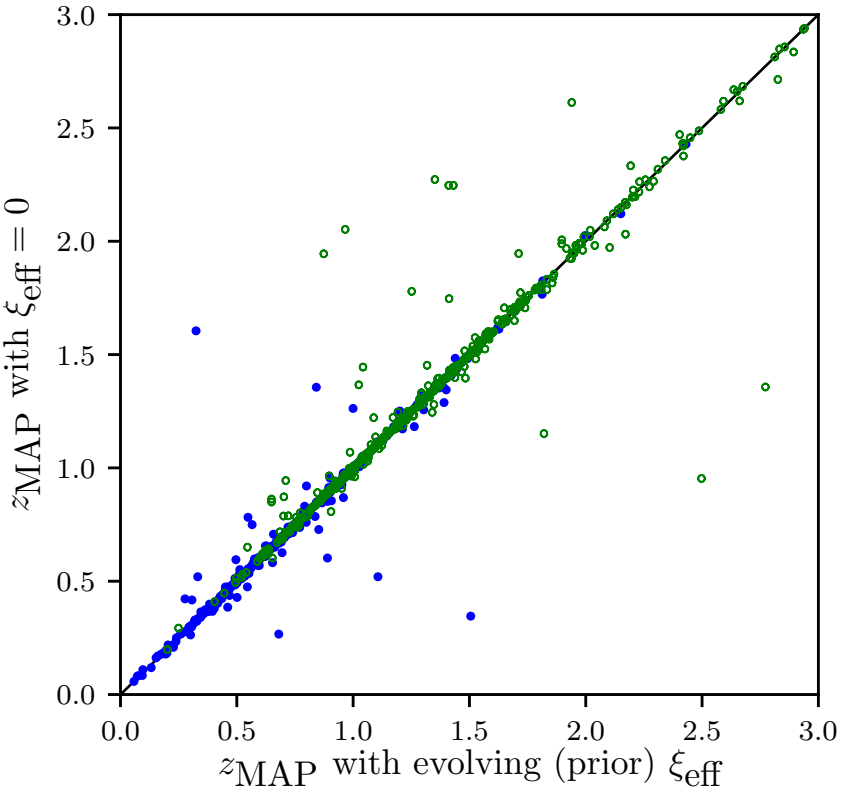

Figure 6.2: Comparison of the maximum *a posteriori* point estimates including the effective correlation function and neglecting it, for sources simulated from a prior that includes it. The lower redshift constituents z_α are plotted with closed blue markers, and z_β are plotted with open green markers. Most sources show negligible differences, while sources that show large differences are multimodal. In these sources, small differences in the posterior result in point estimates moving between modes of slightly different heights, illustrating a limitation of point estimates.

in Figure 6.2. The vast majority of sources show negligible differences, and visually inspecting the posteriors with larger changes shows these are highly multimodal, with modes of comparable heights. In these cases, small differences in the posteriors result in larger differences in point estimates as the maximum *a posteriori* value moves between modes. This is a limitation of point estimates, and can be mitigated by using the full information content of the posterior distributions, which do not vary strongly.

Due to the small effect, our results throughout include a simple non-evolving correlation function with $r_0 = 5\,\mathrm{Mpc}\,h^{-1}$ and $\gamma = 1.77$ (Peebles, 2001). A plot of this is given in Figure 6.3.

Inserting the correlation function allows us to write the joint redshift prior sep-

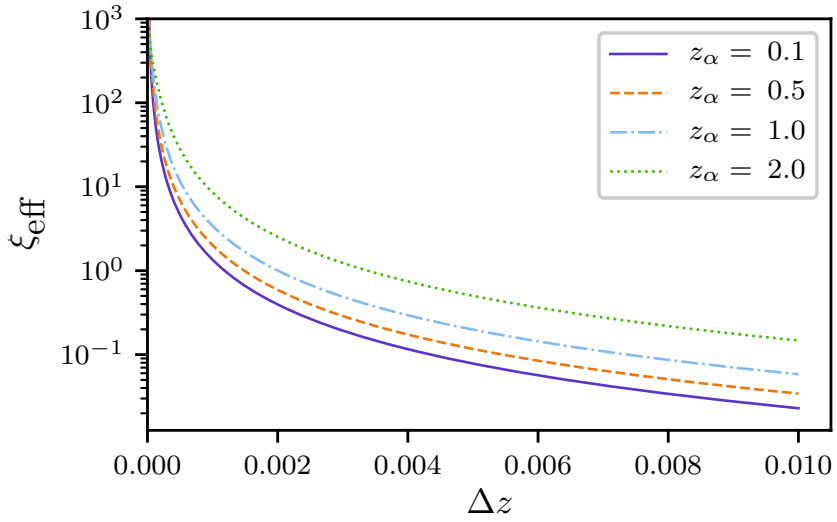

Figure 6.3: Plot of the effective correlation function ξ_{eff} vs $\Delta z \equiv z_\beta - z_\alpha$ for various z_α used for the results throughout.

arated by constituent as

$$P\Big(\{z\}\,\Big|\,\{t\},\{m_0\},\chi,N\Big) = \big[1+\xi_\chi^{(N)}(\{z\})\big]\prod_{\alpha=1}^{N} P\Big(z_\alpha\,\Big|\,t_\alpha,m_{0,\alpha}\Big)\,. \tag{6.22}$$

We separate the joint magnitude prior by assuming that the only correlation between the constituent magnitudes is from the effect of a selection function $S\left(\{m_0\}\right)$ applied to the total magnitude, as discussed in section 6.1.4. The magnitude prior can then be written as

$$P\Big(\{m_0\}\,\Big|\,N\Big) = S\Big(\{m_0\}\Big)\prod_{\alpha=1}^{N} P\Big(m_{0,\alpha}\Big)\,, \tag{6.23}$$

Finally, we impose a sorting condition. Without this, the constituents would be exchangeable, i.e., swapping the constituent labels $\alpha, \beta \ldots$ would have no effect on the prediction of the model. As a result, the marginalised posterior for the redshift of a single constituent would contain contributions from every constituent in the source, as demonstrated in Figure 6.4.

Imposing a sorting condition on either the magnitudes or the redshifts would have the same effect of breaking the exchangeability of the constituents. In our tests, sorting by redshift produced posteriors that recovered the true redshift more successfully. However, in high redshift samples, there is an intrinsic colour degeneracy that can occasionally cause problems with a redshift sorting condition.

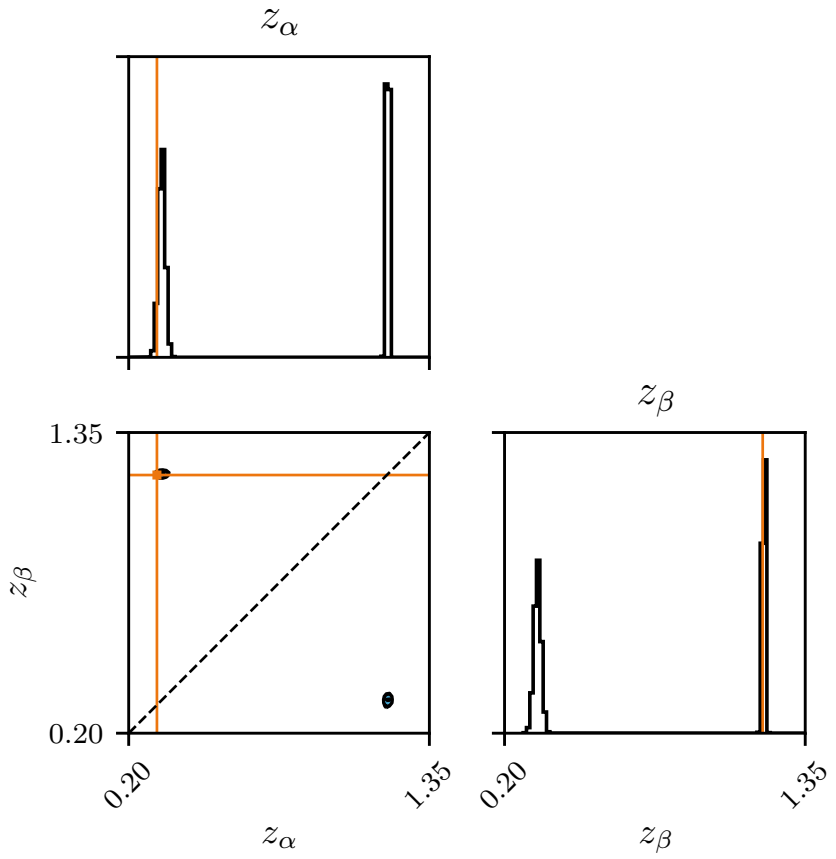

Figure 6.4: By not imposing a sorting condition, the constituents in a source are exchangeable. This is demonstrated here for a simple two-constituent blend with redshifts $z_\alpha = 0.31$, $z_\beta = 1.19$ as indicated by the orange lines. As a result of the exchangeability, the 2D marginal redshift distribution is symmetric about the dashed black line, and each 1D posterior contains a distinct peak for each constituent.

The Lyman break and Balmer break are absorption features occurring at 912Å and 3650Å respectively. If photometry over a sufficiently wide wavelength range is not available, a Lyman break at high redshift can be confused with a Balmer break at low redshift (e.g., Graham et al., 2018). If the sample is deep enough that these high redshift solutions are not unlikely *a priori*, this can cause bimodal posteriors and contribute to catastrophic outliers (Brimioulle et al., 2008).

Consider the case of a two-constituent blend where the redshift of one constituent is well constrained but the other has a bimodal posterior. If the well constrained redshift happens to lie between these two modes, it will appear in the 1D marginal distributions of each constituent redshift, as whether it is the lower or higher redshift object depends on which of the two degenerate peaks is being sampled. In this case, sorting by magnitudes would result in a posterior more representative of the underlying system, where the redshift of one constituent is well constrained while the other has two well separated modes. We did not find this to be a problem in our tests however, and so apply redshift sorting throughout.

The sorting condition Λ_α is imposed by introducing Heaviside step functions Θ into the product over constituents, and is defined as

$$\begin{aligned} \Lambda_\alpha &= 1 \quad \text{for} \quad \alpha = 1 \\ &= \Theta(q_{\alpha-1} - q_\alpha) \quad \text{otherwise,} \end{aligned} \tag{6.24}$$

where q is either z or m_0 depending on whether redshift or magnitude sorting is used. In summary, the posterior for the fully-blended case is given by

$$P\Big(\{z\},\{m_0\}\,\Big|\,\hat{\boldsymbol{F}},\hat{F}_0,\chi,N\Big) \propto \sum_{i=1}^{T^N} P\Big(\hat{\boldsymbol{F}}\,\Big|\,\{z\},\{t\}_i,\{m_0\},N\Big)P\Big(\hat{F}_0\,\Big|\,\{m_0\}\Big)\left[1+\xi_\chi^{(N)}(\{z\})\right] \times S\Big(\{m_0\}\Big)\prod_{\alpha=1}^{N}\Lambda_\alpha P\Big(z_\alpha\,\Big|\,t_\alpha,m_{0,\alpha}\Big)P\Big(t_\alpha\,\Big|\,m_{0,\alpha}\Big)P\Big(m_{0,\alpha}\Big)\,. \tag{6.25}$$

6.1.4 Accounting for selection effects

When considering the total apparent magnitude of a source, we must account for the selection effect of the survey observing it. Galaxy surveys typically select sources by imposing cuts on the apparent magnitude they observe $m < m_{\rm lim}$ since they cannot observe arbitrarily faint sources. As we are sampling *intrinsic* magnitudes rather than *observed* magnitudes, these selection effects do not impose a hard cut in our magnitude prior.

Consider a source with an intrinsic apparent magnitude exactly equal to the survey magnitude limit. Assuming a normal distribution for the observational error, the probability of observing this source is 1/2, since its observed apparent magnitude is equally likely to have been scattered above and below the magnitude cut. However, since objects in the sample have been detected by definition, we know the source must have been scattered brighter, effectively breaking the symmetry of the error distribution. As a result, intrinsic apparent magnitudes around the magnitude limit are less probable and should be downweighted.

To account for this, we follow the approach described in Leistedt et al. (2016) for including a selection effect. A discrete variable D representing the fact that an object was detected is introduced, and each term in the posterior is conditioned on it. We assume that our selection effect is imposed on a single selection band. Without loss of generality, we derive the effect by assuming that the selection band is the reference band b_0 and so only the reference band likelihood is affected. Conditioning on D, the likelihood can be written using Bayes rule as

$$P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N, D\Big) = \frac{P\Big(D \,\Big|\, \hat{F}_0, \{m_0\}, N\Big) P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big)}{\int_0^\infty P\Big(D \,\Big|\, \hat{F}_0, \{m_0\}, N\Big) P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big) \mathrm{d}\hat{F}_0} \,. \tag{6.26}$$

The numerator of equation 6.26 is equal to the likelihood defined in equation 6.8 since the probability of detection for an object that we know has been observed is $P\Big(D \,\Big|\, \hat{F}_0, \{m_0\}, N, \Big) = 1$. After integrating over $\hat{F}_0$, the denominator depends only on $\{m_0\}$ and represents the effect of the magnitude selection. We therefore choose to write this term as part of the joint magnitude prior, defining the selection effect

$$S\Big(\{m_0\}\Big) = \int_0^\infty P\Big(D \,\Big|\, \hat{F}_0, \{m_0\}, N, \Big) P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big) \mathrm{d}\hat{F}_0 \tag{6.27}$$

that appears in the posterior in equation 6.25. The selection is a hard cut based on the observed flux, and so

$$\begin{aligned} P\Big(D \,\Big|\, \hat{F}_0, \{m_0\}, N, \Big) &= 1 \quad \text{for} \quad \hat{F}_0 > 10^{-0.4 m_{\mathrm{lim}}} \\ &= 0 \quad \text{otherwise.} \end{aligned} \tag{6.28}$$

Thus, the integral becomes

$$S\Big(\{m_0\}\Big) = \int_{10^{-0.4 m_{\mathrm{lim}}}}^\infty P\Big(\hat{F}_0 \,\Big|\, \{m_0\}, N\Big) \mathrm{d}\hat{F}_0 \,. \tag{6.29}$$

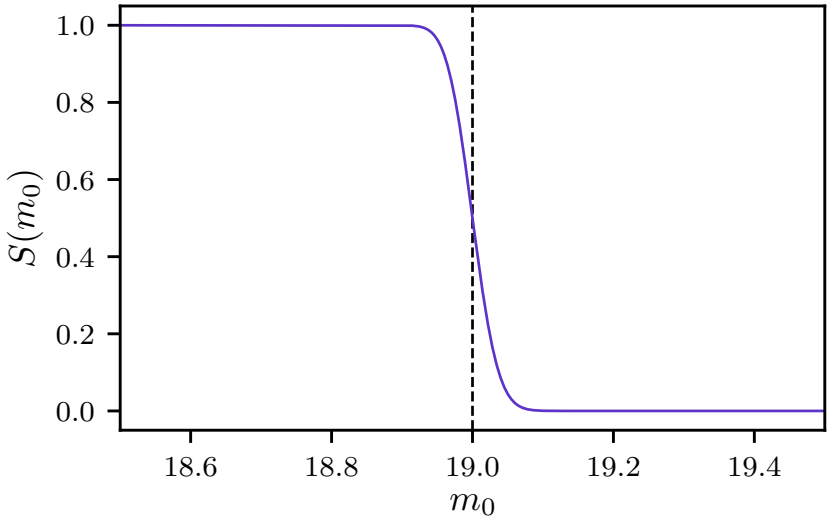

Figure 6.5: Plot of the selection function for a typical source from the GAMA blended sources catalogue used in section 6.5. The dashed line shows the magnitude limit for this source $m_{\rm lim} < 19$.

Since the reference band likelihood is assumed Gaussian, this can be written in terms of the normal cumulative distribution function as $S\left(\{m_0\}\right) = 1 - \Phi(\hat{F}_0)$, where Φ is defined for a Gaussian distribution with mean μ and standard deviation σ to be

$$\Phi(x) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] . \tag{6.30}$$

Inserting this into equation 6.29, the effect of the magnitude selection can be written as

$$S\Big(\{m_0\}\Big) = \frac{1}{2} - \frac{1}{2}\,\mathrm{erf}\left(\frac{10^{-0.4m_{\rm lim}} - \sum_{\alpha=1}^{N} 10^{-0.4m_{0,\alpha}}}{\sigma_0\sqrt{2}}\right) . \tag{6.31}$$

By replacing the reference-band flux $\sum_{\alpha=1}^{N} 10^{-0.4m_{0,\alpha}}$ with the model flux $F^{(N)}_{\{t\},b}\big(\{z\},\{m_0\}\big)$, this selection can performed on any band. The selection function would then also be dependent on the redshifts and templates, i.e., $S\Big(\{z\},\{t\},\{m_0\}\Big)$. This choice of selection band is included in the implementation described in section 6.3.3.

A plot of this selection function for a galaxy from the GAMA blended sources catalogue, described in section 6.5, is shown in Figure 6.5.

6.1.5 Specifying the priors

Like all Bayesian methods, the choice of priors should be problem dependent. For ease of comparison, we use the parametric forms given by Benítez (2000) with an additional magnitude prior. However, we stress that this choice is not a necessary one for our method and any joint $P\Big(z, t, m_0\Big)$ prior may be used.

The Benítez (2000) template and redshift priors are given by

$$P\Big(t \,\Big|\, m_0\Big) = f_t \mathrm{e}^{-k_t(m_0 - m_{\min})} \tag{6.32}$$

and

$$P\Big(z \,\Big|\, t, m_0\Big) \propto z^{\alpha_t} \exp\left\{-\left[\frac{z}{z_{0,t} + k_{\mathrm{m},t}(m_0 - m_{\min})}\right]^{\alpha_t}\right\} \tag{6.33}$$

respectively, where $m_{\min}$ is the bright-end magnitude cut as described below. The parameters α_t, $z_{0,t}$, $k_{\mathrm{m},t}$, f_t and k_t are set separately for early, late and irregular template types. Their values are found using the procedure discussed in section 6.1.6 and are listed in Table 6.2.

We use a magnitude prior given by

$$P\Big(m_0\Big) \propto 10^{\,\phi m_0}\,. \tag{6.34}$$

The value $\phi = 0.6$ gives the expression for the expected galaxy number counts in a homogeneous, Euclidean universe (Yasuda et al., 2001), though we leave ϕ free to also be found using the procedure discussed in section 6.1.6. When fitted to the GAMA blended sources catalogue (Holwerda et al., 2015) discussed in section 6.5, this value was found to be $\phi = 0.705$, though the difference in results compared to fixing $\phi = 0.6$ was negligible, and fitting to other datasets may yield a different value.

Since the selection effect applies to the total source flux, individual constituents may be fainter than the survey magnitude limit, and so unobservable outside of a blended source. As a result, an analytic magnitude prior is required to describe the distribution of constituent magnitudes so that it can be used at faint magnitudes, where observations of individual constituents are unavailable.

For the reasons discussed in section 6.3.2, we also apply a hard minimum and maximum cut to each constituent redshift z_α and constituent magnitude $m_{0,\alpha}$. This cut has little effect on the redshift priors which already go towards zero at large redshift; the same is true of the magnitude prior at bright magnitudes. The faint-end of the magnitude prior of the brightest constituent is also already forced towards zero by the selection function. This is because in an N-constituent blend, the flux of the brightest

constituent must be at least $1/N$ that of the total source flux, by definition. The magnitudes of the other constituents are not constrained in this way however, and so this cut represents a sharp boundary in the prior.

In our tests, the results of the redshift estimation were not strongly dependent on the position of this faint-end magnitude cut $m_{\max}$. However, the evidence calculation described in section 6.3.1 *is* dependent on its position, as changing the position of the cut alters the prior volume integrated over in equation 6.42. As a result, the position of this cut must be decided; it defines the limit where a galaxy is considered to contribute to a blend, and is therefore problem-dependent.

In principle, one could consider a galaxy to be blended if another arbitrarily faint galaxy lies along the same line of sight. In practice however, observations have limited precision, and the flux of an extremely dim galaxy cannot be detected. In other words, a sufficiently dim galaxy should no longer be considered a blended constituent, but rather a contribution to the noise.

In practice, a simple method to set this cut is to fix it for the entire sample. However, the argument above suggests that this cut should be dependent on the noise of the observation, i.e., that $m_{\max}$ should be set to the faintest magnitude that would have an observable effect. Fixing $m_{\max}$ is effectively an assumption that the sample has sufficiently homogeneous noise properties that the change in this faintest magnitude is negligible. For a sample where this is not the case, the magnitude cut can be set as an $n\sigma_0$ flux deviation, i.e.,

$$m_{\max} = -2.5 \log_{10}(n\sigma_0)\,, \tag{6.35}$$

where σ_0 is the error on the reference band flux which varies for each source. In the tests in section 6.5, we test both of these methods of setting $m_{\max}$.

6.1.6 Calibrating the priors using spectroscopic information

The joint prior is conditioned on a set of parameters θ, i.e., $P\Big(z, t, m_0 \,\Big|\, \theta\Big)$, the posterior distribution of which we wish to infer. We can use spectroscopic information of a sample of galaxies from the population of interest to calibrate the above priors as suggested by Benítez (2000).

We assume here that this calibration is done with unblended galaxies, though this procedure can be extended to include blended galaxies too, provided that the number of constituents N is known *a priori*. In that case, the reference band magnitudes of each constituent would need to be included as a parameter in this model, and either sampled along with θ or marginalised out of the posterior analytically.

We consider a sample of G galaxies with photometry and spectroscopic redshifts $\hat{z}_s$. These redshifts are assumed to be exact, i.e., we neglect the error on $\hat{z}_s$. The set notation here now runs over each independently observed galaxy, not the blended constituents as before.

We start by writing this posterior as a marginalisation over the photometric redshift model parameters for each galaxy and applying Bayes rule. Since the likelihood is independent of the prior parameters, we condition on θ in the prior only, giving

$$P\Big(\theta \,\Big|\, \{\hat{z}_s\}, \{\hat{\boldsymbol{F}}\}, \{\hat{F}_0\}\Big) \propto \int \mathrm{d}^G\{z\} \int \mathrm{d}^G\{m_0\} \times \sum_{i=1}^{T^G} P\Big(\{\hat{z}_s\}, \{\hat{\boldsymbol{F}}\}, \{\hat{F}_0\} \,\Big|\, \{z\}, \{t\}_i, \{m_0\}\Big) P\Big(\theta, \{z\}, \{t\}_i, \{m_0\}\Big) \,. \tag{6.36}$$

We apply product rule to separate the joint prior and remove other unnecessary conditioning. We also assume that the galaxies in the sample are independent, and so all terms not shared across the population (i.e., $P(\theta)$) can be written as a product over galaxies. The posterior then becomes

$$\begin{aligned} P\Big(\theta \,\Big|\, \{\hat{z}_s\}, \{\hat{\boldsymbol{F}}\}, \{\hat{F}_0\}\Big) \propto & P\Big(\theta\Big) \prod_{g=1}^{G} \int \mathrm{d}z_g \int \mathrm{d}m_{0,g} \times \\ & \sum_{i_g=1}^{T} P\Big(\hat{F}_{0,g} \,\Big|\, m_{0,g}\Big) P\Big(\hat{\boldsymbol{F}}_g \,\Big|\, z_g, t_{i_g}, m_{0,g}\Big) \times \\ & P\Big(\hat{z}_{s,g} \,\Big|\, z_g\Big) P\Big(z_g, t_{i,g}, m_{0,g} \,\Big|\, \theta\Big) \,. \end{aligned} \tag{6.37}$$

By assuming that the spectroscopic redshifts are exact, the redshift likelihood can be written as a delta function, i.e., $P\Big(\hat{z}_{s,g} \,\Big|\, z_g\Big) = \delta\Big(z_g - \hat{z}_{s,g}\Big)$. We also assume that the error on the reference band magnitude is negligible, allowing us to write $P\Big(\hat{F}_{0,g} \,\Big|\, m_{0,g}\Big) = \delta\Big(m_{0,g} - \hat{m}_{0,g}\Big)$, where $\hat{m}_{0,g} = -2.5 \log_{10}\Big(\hat{F}_{0,g}\Big)$ is the reference band flux of galaxy g, converted to magnitudes. Replacing these likelihoods with delta functions, the marginalisation can be done analytically using the sifting property of the delta function to give

$$P\Big(\theta \,\Big|\, \{\hat{z}_s\}, \{\hat{\boldsymbol{F}}\}, \{\hat{F}_0\}\Big) \propto P\Big(\theta\Big) \prod_{g=1}^{G} \sum_{i_g=1}^{T} P\Big(\hat{\boldsymbol{F}}_g \,\Big|\, \hat{z}_{s,g}, t_{i_g}, \hat{m}_{0,g}\Big) P\Big(\hat{z}_{s,g}, t_{i_g}, \hat{m}_{0,g} \,\Big|\, \theta\Big) \,. \tag{6.38}$$

To find the prior parameters θ that maximise this posterior, we use L-BFGS-B (Byrd et al., 1995), a local optimisation algorithm that approximates the Hessian of the objective function and optimises the parameters subject to simple box constraints;

Table 6.2: The maximum *a posteriori* values of the prior parameters for the GAMA blended sources catalogue found after calibrating using 26782 unblended sources.

Parameters	Early	Late	Irregular	Type-independent
α_t	1.59	1.53	1.30	-
$z_{0,t}$	0.016	0.019	0.066	-
$k_{m,t}$	0.048	0.048	0.022	-
k_t	0.044	0.024	-	-
f_t	0.45	0.51	-	-
ϕ	-	-	-	0.71

we use these constraints to ensure our parameters are positive. This method requires first-order derivatives which we approximate through a finite difference method.

The result of this procedure is an estimate of the maximum *a posteriori* values of the prior parameters. Throughout this chapter, we use these values in the priors directly. In principle, these parameters could form part of a hierarchical model and be marginalised out as nuisance parameters. However, this would significantly increase the dimensionality of the parameter space to be sampled and, thus, the computation time required for each source. Table 6.2 lists the values of these prior parameters GAMA test described in section 6.5. A plot of samples drawn from the resulting prior is shown in Figure 6.6.

6.2 Partially-blended sources

We can modify the formalism above for the case of sources for which every constituent does not contribute to every observation. We refer to these as partially blended sources. This can be the case when combining photometry from a wide range of wavelengths, e.g., optical and far-infrared observations. This partial blending may also occur for some sources observed in both a ground-based and space-based survey, as the latter does not suffer from atmospheric seeing and so can achieve a higher spatial resolution. An example of a pair of such surveys is LSST (Ivezić et al., 2019) and Euclid (Laureijs et al., 2011). Utilising resolved photometry from Euclid could improve the precision of photometric redshifts of sources that are blended in the higher signal-to-noise observations of LSST. This possibility is explored using simulated observations in section 6.4.2.

To generalise the method for this case, we introduce the measurement-constituent mapping $\delta_{\alpha,m}$, an $N \times N_m$ matrix, where N_m is the number of measurements, a generali-

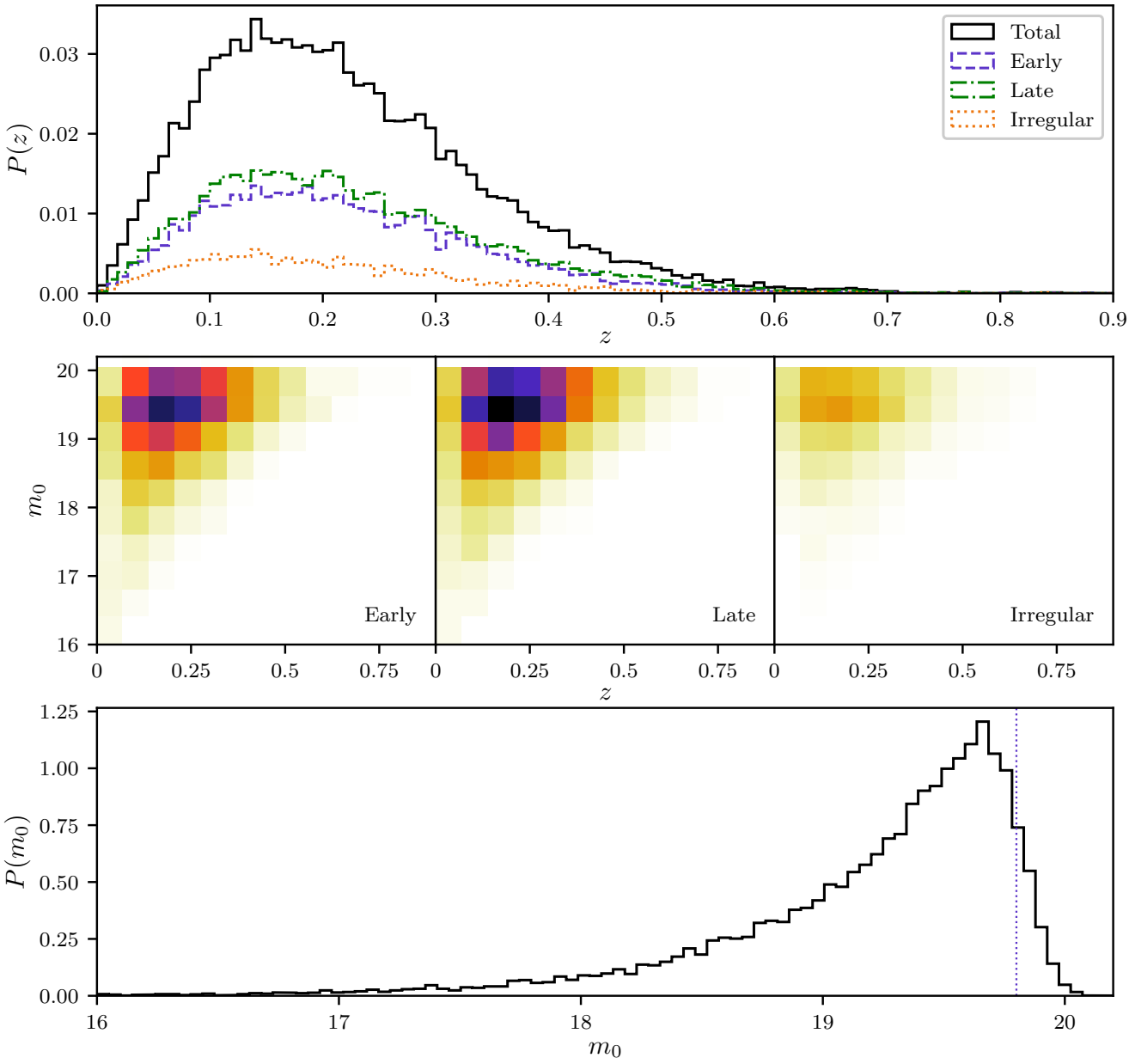

Figure 6.6: Plot of the prior found for the test on the GAMA blended sources catalogue after calibrating using 26782 unblended sources. The dashed line in the bottom panel shows a magnitude limit of $r < 19.8$.

sation of the number of bands in the fully-blended case. This measurement-constituent mapping acts as an indicator variable, consisting only of zeros and ones indicating whether a particular constituent is present in a particular measurement.

An example of such a matrix is given below. Consider data containing $N_m = 6$ photometric measurements of $N = 2$ constituents. The first four measurements are of individually resolved constituents, while the final two measurements are blended. In a typical use case, we might expect the resolved measurements of each constituent to share filter bands, though the model does not require this. In this example, the measurement-constituent mapping is given by

$$\delta = \begin{bmatrix} 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \end{bmatrix} . \tag{6.39}$$

We can then write the blended flux of N constituents at a redshift z in measurement m as

$$F^{(N)}_{\{t\},m,\delta}\big(\{z\},\{m_0\}\big) = \sum_{\alpha=1}^{N} \delta_{\alpha,m} \frac{10^{-0.4 m_{0,\alpha}}}{T_{t_\alpha,b_0}\big(z_\alpha\big)} T_{t_\alpha,m}\big(z_\alpha\big) . \tag{6.40}$$

The only modification to the posterior of the fully-blended case needed to accommodate the partial-blending is to the sorting condition. As described in section 6.1.3, the purpose of this condition is to prevent the exchangeability of constituents However, this is not necessary in the partially blended case. Here, the constituents are intrinsically different as they appear individually in separate measurements and so are not exchangeable. As a result, we drop the sorting condition for the partially blended case, i.e., $\Lambda_\alpha = 1$ over the entire parameter space. The posterior for the partially blended case is then given by

$$\begin{aligned} P\Big(\{z\},\{m_0\} \,\Big|\, \hat{\boldsymbol{F}},\hat{F}_0,\chi,N,\delta\Big) \propto \sum_{i=1}^{T^N} P\Big(\hat{\boldsymbol{F}} \,\Big|\, \{z\},\{t\}_i,\{m_0\},N,\delta\Big) P\Big(\hat{F}_0 \,\Big|\, \{m_0\}\Big) \times \\ \big[1+\xi^{(N)}_\chi\big(\{z\}\big)\big]\, S\Big(\{m_0\}\Big) \prod_{\alpha=1}^{N} P\Big(z_\alpha \,\Big|\, t_\alpha, m_{0,\alpha}\Big) \times \\ P\Big(t_\alpha \,\Big|\, m_{0,\alpha}\Big) P\Big(m_{0,\alpha}\Big) . \end{aligned} \tag{6.41}$$

6.3 Inference using Nested Sampling

6.3.1 Determining the number of constituents with model comparison

The posteriors in equations 6.25 and 6.41 are defined for a specific number of constituents N. In general however, this number of constituents is not known *a priori*. We therefore need a method to determine how many constituents are present in a source. Since our model is defined for a fixed number of constituents, we treat finding the number of constituents in a source as a model comparison problem.

Bayesian model comparison involves the calculation of the evidence $\mathcal{Z}$, an integral over the product of the prior and the likelihood (e.g., Trotta, 2008). Given a data vector $\boldsymbol{d}$, a model m and a set of model parameters $\{\theta\}$, the evidence is defined as

$$\mathcal{Z} \equiv P\Big(\boldsymbol{d}\,\Big|\,m\Big) = \int P\Big(\boldsymbol{d}\,\Big|\,\{\theta\}, m\Big) P\Big(\{\theta\}\,\Big|\,m\Big) \mathrm{d}\{\theta\}\,. \tag{6.42}$$

This evidence term plays the role of the normalisation of the posterior and so is typically ignored in parameter inference problems where this normalisation is irrelevant. However, the evidence is the quantity of interest for model comparison problems. The ratio of the posterior probabilities of two models is proportional to the ratio of their evidences, a quantity known as the Bayes factor. By considering the number of constituents in a source as the model, we can write the relative probability of the source containing n constituents compared to m constituents as

$$\mathcal{P}_{n,m} = \frac{P\Big(N = n\,\Big|\,\hat{\boldsymbol{F}}, \hat{F}_0\Big)}{P\Big(N = m\,\Big|\,\hat{\boldsymbol{F}}, \hat{F}_0\Big)} = \frac{P\Big(\hat{\boldsymbol{F}}, \hat{F}_0\,\Big|\,N = n\Big)}{P\Big(\hat{\boldsymbol{F}}, \hat{F}_0\,\Big|\,N = m\Big)} \frac{P\Big(N = n\Big)}{P\Big(N = m\Big)}\,. \tag{6.43}$$

Considering the cases of either isolated galaxies or blends of two constituents, the model prior ratio $P\big(N = 2\big)/P\big(N = 1\big)$ represents the probability that a galaxy will be blended. Dawson and Schneider (2014) estimate the number of sources observed by LSST that will be blended by convolving Hubble Space Telescope images with a Gaussian point spread function (PSF) like that of LSST. They found this number to be $45 - 55\%$ of the total sources observed, with $15 - 20\%$ of observed sources classified as catastrophic blends that would be identified as single sources by fitting a profile template to a galaxy image. Chang et al. (2013) estimates that the rejection of blended sources will reduce the number density of LSST sources by 16%, though this estimate does not include the catastrophic blends of above. Studies such as these

using existing high-resolution data or simulated observations can inform the blending prior ratio. Throughout this chapter, we present results where this prior ratio is $P\big(N=2\big)/P\big(N=1\big)=1$, i.e., we do not prefer either of the blended or single-constituent models *a priori*, though this information can be trivially included.

6.3.2 Nested sampling using MultiNest

Calculating the evidence directly through numerical integration presents a difficult technical problem, particularly as the number of dimensions increases. To avoid this, we use Nested sampling (Skilling, 2006), a Monte Carlo method for estimating the evidence while also sampling the posterior for parameter inference. Nested sampling reduces the problem of estimating the evidence to sampling a series of increasing likelihood thresholds, i.e., progressively smaller prior volumes nested within one another. Equation 6.42 can then be calculated using a one-dimensional quadrature integration method over this prior volume.

The computationally difficult part of the nested sampling algorithm is sampling a new point from within the potentially complicated boundary defined by the likelihood threshold. The MultiNest sampler (Feroz et al., 2009) does this efficiently by sampling from a collection of ellipses approximating this boundary rather than the prior itself. This collection of ellipses is formed by performing a clustering analysis on a fixed-sized set of the previous samples, known as the live points. A new sample is drawn from these ellipses, replacing the lowest likelihood point which is removed and stored as a posterior sample. Samples are rejected until the likelihood boundary is respected, though this occurs less frequently than when naively rejection sampling the prior.

The use of multiple ellipses when sampling has another distinct advantage in that it naturally enables efficient sampling of multimodal posteriors, since each mode is assigned a separate ellipse while low probability regions between these modes are avoided. Multimodality is a feature that can cause difficulties for MCMC samplers, as moving from one mode to another requires a move across the low probability region separating them. As a result, these samplers can fail to explore the full posterior distribution, instead sampling only a single mode. We expect our problem to exhibit this multimodal behaviour due to the degeneracies described in section 6.1.3, and so require a sampling method suited to this case.

The need for nested sampling methods to sample from the prior imposes some constraints on our choice of prior. MultiNest natively samples from a unit side-length hypercube and these samples are transformed into samples of the prior using a prior

transform function. However, due to the discrete marginalisation over template, we cannot separate the posterior to define a prior transform function. As a result, we take the approach suggested by Feroz et al. (2009) of defining a uniform prior to sample from, and defining the 'likelihood' for MultiNest as our marginalised posterior.

This has two main effects. Firstly, the sampling is likely to be less efficient, as the prior sampling step is not guided by the true prior, and so low-prior regions may be sampled frequently. Secondly, sampling from a uniform prior necessitates imposing a hard cut on the prior range of each parameter. Since the location of these cuts effects the value of the evidence $\mathcal{Z}$, they should not be imposed thoughtlessly. At high redshift and bright magnitudes, the priors tend to zero, meaning that the exact positions of these cuts have negligible effect on the evidence. However, this is not the case for the faint-end of the magnitude priors; setting this cut is discussed is section 6.1.5.

6.3.3 `blendz` package

We have written a Python package `blendz` to perform the redshift inference of blended sources described in sections 6.1 and 6.2, and the identification of the number of constituents using model comparison described in section 6.3.1. The package supports analysis of blends with an arbitrary number of constituents using either the included or user-supplied template sets. The output of such an analysis is a set of samples from the joint posterior for each number of constituents considered, and an estimate of the Bayes factor for model comparison. The model comparison can then easily include a model prior through multiplication of the Bayes factor.

The package is also written in an object-orientated way, allowing the user to easily redefine the priors. While the supplied prior is used in this work with galaxies of either early, late or irregular types, it is written to be calibrated and used with any number of possible types. For blended sources of more then two constituents, the excess probability term $\xi_\chi^{(N)}$ is defined recursively to use the correct combination of two-point terms and assumes higher order correlations are negligible.

Documentation and instructions for installation can be found at `http://blendz.readthedocs.io`. The package can also be immediately installed from the official Python Package Index[2] by using the `pip install blendz` command. Finally, the source is available in a `git` repository hosted at `https://github.com/danmichaeljones/blendz`.

[2] `https://pypi.org/`

6.4 Results from mock observations

6.4.1 Fully-blended sources

As an initial test of the method, we used a Monte Carlo simulation to create a set of mock photometric observations to test our method against. These mock observations simulate an optical survey using the six LSST optical filters u, g, r, i, z, Y (LSST Science Collaboration et al., 2009), with an r-band magnitude selection of $m_{\rm lim} = 24$. We also applied hard cuts to the constituent magnitudes of $m_{\rm min} = 19$ and $m_{\rm max} = 26$. We then generated 1000 sources, each of which is a blend of two constituents in all bands. This was done by sampling a prior describing this distribution of objects using the Markov Chain Monte Carlo (MCMC) sampler `emcee` (Foreman-Mackey et al., 2013) to generate the true parameters $\{z\}, \{t\}, \{m_0\}$ for each simulated source. A plot of this prior distribution, plotted using `corner.py` (Foreman-Mackey, 2016), is shown in Figure 6.7.

The effect of the selection function and the faint-end magnitude cut can be seen clearly in the two-peaked shape of the marginal distributions of $m_{0,\alpha}$ and $m_{0,\beta}$. The brighter-magnitude peak is a result of the selection function. In the single-constituent case, this would cause the prior to tend to zero at faint magnitudes. In the two-constituent case however, the magnitude priors of each constituent extend beyond $m_{\rm lim}$ as the selection effect is applied to the combined magnitude of both constituents. The brighter constituent in a two-constituent blend must, by definition, contribute at least half of the total flux. As a result, the selection effect prevents the magnitude of this constituent from being too faint. Since we impose the sorting condition on redshifts, and the brightest constituent in a source is not exclusively the lower-redshift one, this action of the sorting condition causes the brighter peak in the marginal distributions of both $m_{0,\alpha}$ and $m_{0,\beta}$. If we instead impose the sorting condition on the magnitudes, these distributions become unimodal. Figure 6.7 also shows the effect of the redshift sorting condition in the (z_α, z_β) marginal distribution as a hard diagonal cut.

The model fluxes for these sampled parameters were then generated using the template responses defined in section 6.1.1. We use the template set of Coe et al. (2006), containing one early type, two late type and one irregular type templates from Coleman et al. (1980), two starburst templates from Kinney et al. (1996) and two starburst templates from Coe et al. (2006). This same template set is then used during the inference. This allows a test of the method without the effect of unrepresentative templates, a source of error that is not unique to the case of blended sources.

Finally, we add an observational error to each observation. The flux error in

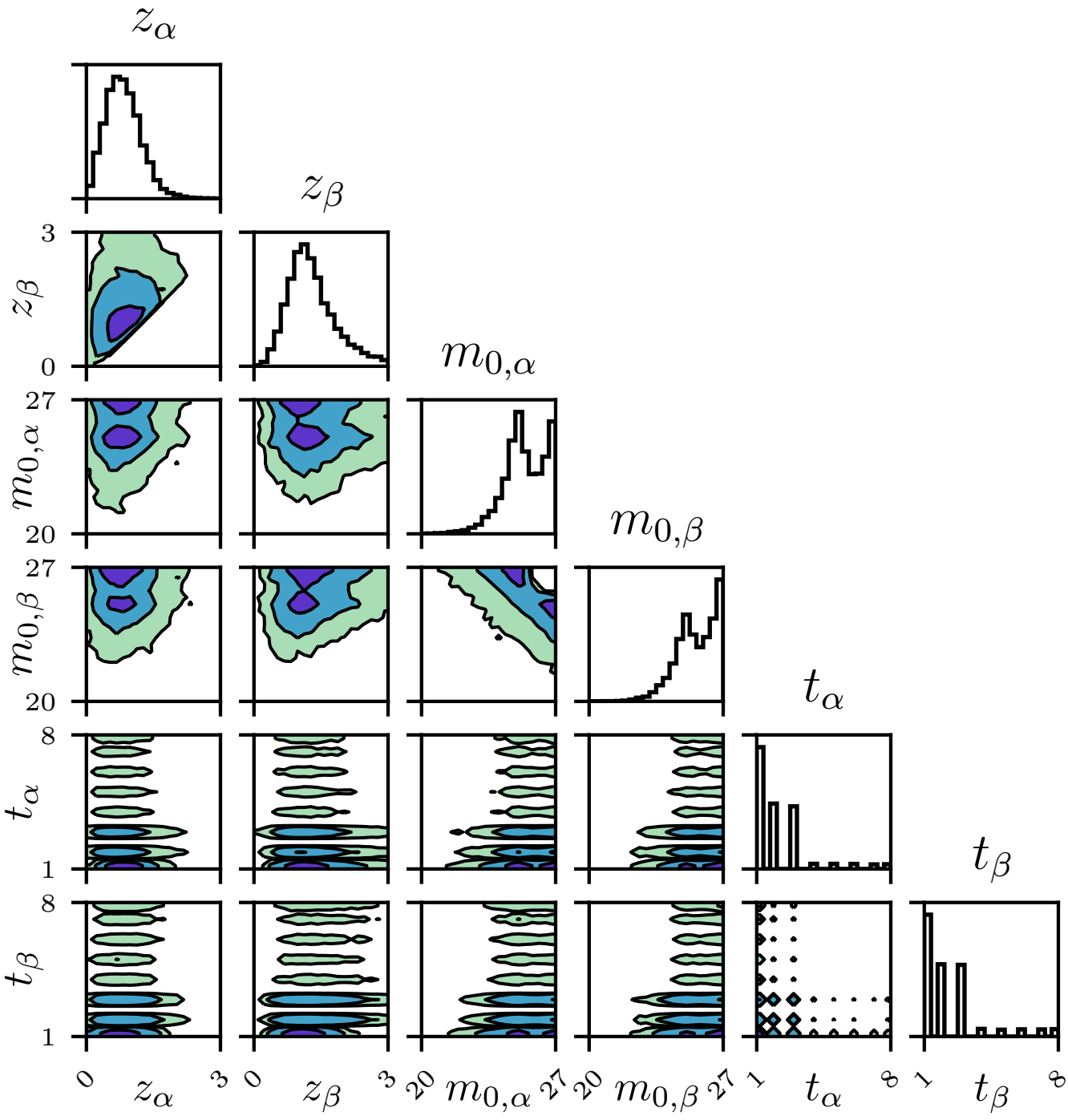

Figure 6.7: Corner plot of the prior sampled to create the mock catalogue. As described in the text, the bimodal shape of the marginal magnitude distributions is a result of both the selection effect and sorting constituents by redshift. The redshift sorting condition can be seen as a hard diagonal cut in the joint redshift distribution.

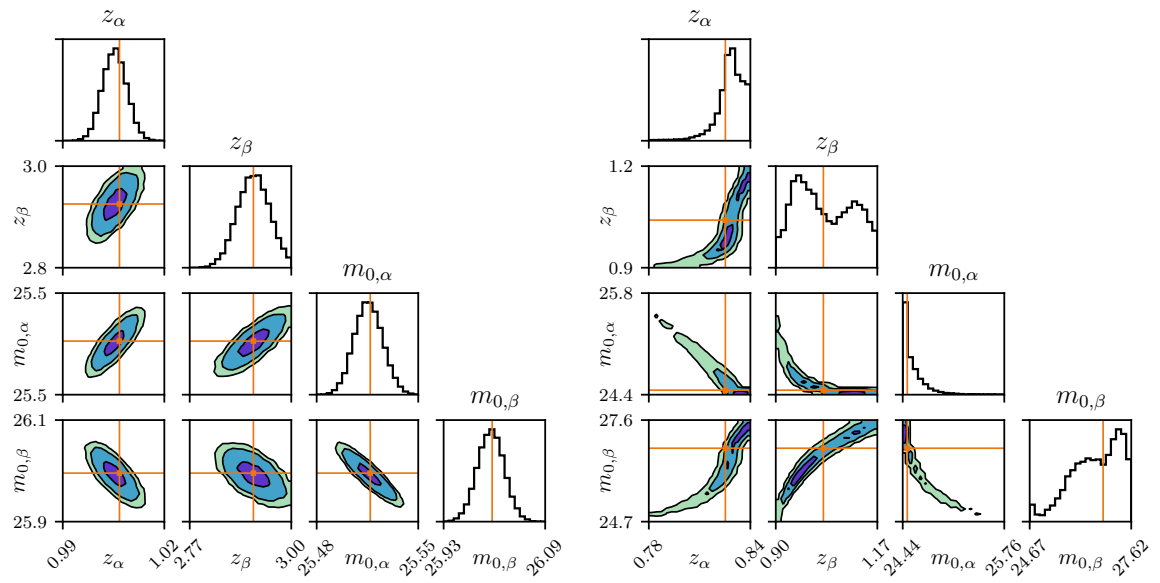

Figure 6.8: The 4D posterior distribution output from our method for two example sources. The true parameter values are shown in orange. The left panel shows a well constrained source with some correlations between constituents, though the true redshift is well recovered. The right panel shows an example of a bimodal posterior that can arise in photometric redshift problems.

band b is randomly drawn from an uncorrelated, zero-centred normal distribution $\sigma_b \sim \mathcal{N}(\sigma_b \,|\, 0, \Sigma)$. The noise is set for all sources to be the final 1σ depth expected from LSST (Ivezić et al., 2019). We use these noisy observations to draw samples from the one- and two-constituent posteriors to test both the redshift determination and model comparison performance, setting the prior to the true distribution the photometry was sampled from.

Figure 6.8 shows two examples of the 4D posterior that is the output from our method for each sample. For plotting purposes, the number of live points used for sampling is larger than that used for the inference and model comparison results throughout this chapter. However, the change in the results is negligible. The left panel shows an example of a well constrained source with a unimodal posterior. This posterior shows correlations between the constituent parameters; this is expected, since the total flux of each band that is well constrained by the observations is split between the constituents. Reducing the model flux in a band of one constituent will result in a compensation in the other constituent, correlating their parameters.

The right panel of Figure 6.8 shows a particularly prominent example of the curved degeneracies that can arise in the blended posteriors. This is due to the total magnitude of the source being well constrained by noisy observations, while constituent

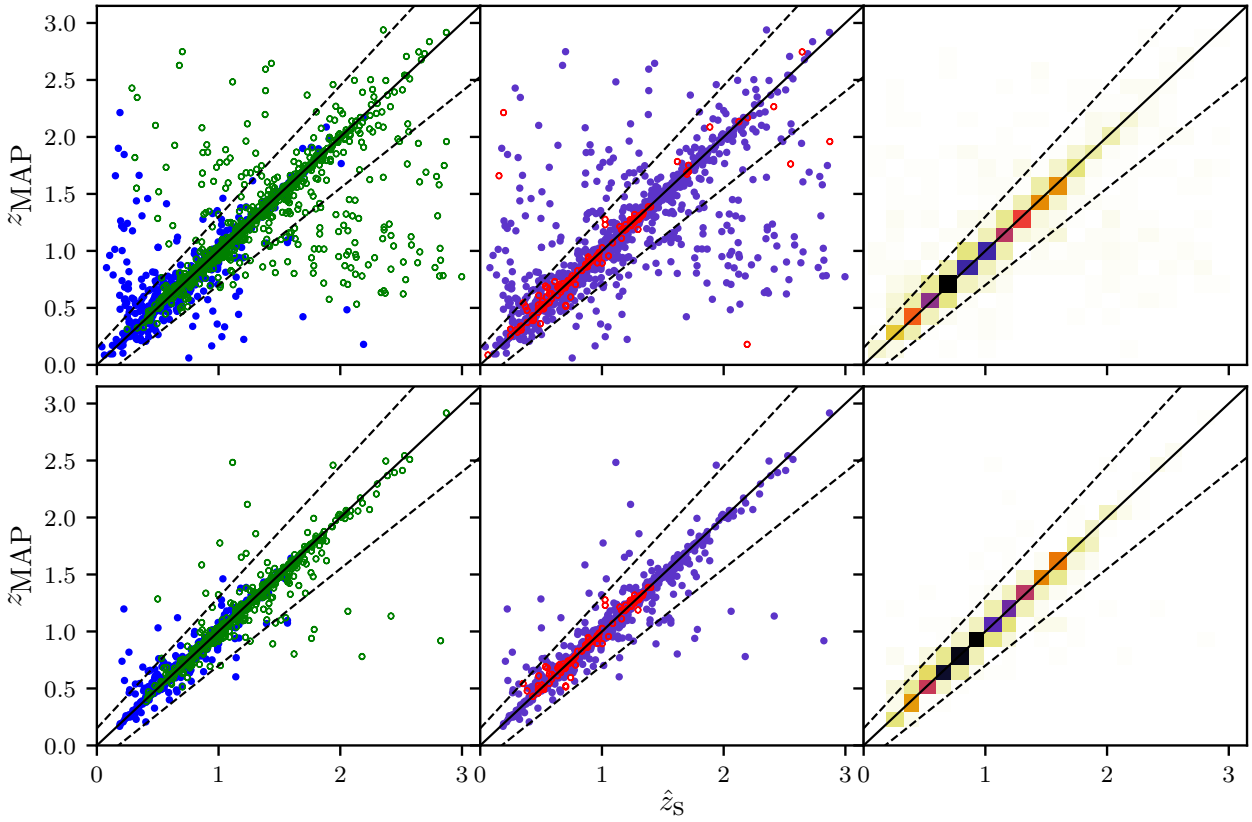

Figure 6.9: Scatter plot comparing the maximum *a posteriori* point estimates from the photometric redshift estimation with the true redshifts for the mock observations. The left panels distinguish the constituents, with z_α plotted with closed blue markers, and z_β plotted with open green markers. The centre panels show the blend identification, with sources identified as blends plotted with closed purple markers, and those misidentified as single sources plotted with open red markers. The right panels show a 2D histogram of the combined sample. Panels in the top row show the results for the full mock catalogue, while the bottom row only includes sources where the standard deviation of samples from each redshift marginal-posterior are sufficiently small, $\sigma_\alpha \leq 0.2 \,\forall\, \alpha$. The dashed lines in each panel show an error of $0.15(1+z)$.

magnitudes are not themselves observable. This leads to a degeneracy that is curved due to the non-linearity of adding magnitudes. A result of this curved degeneracy is bimodality in the marginalised posterior of z_β. However, there is still significant probability density around the true redshift, highlighting the importance of not compressing the information content of a full posterior distribution into only a small set of numbers.

Figure 6.9 shows a comparison of the photometric estimation of the constituent redshifts against their true simulated values. Point estimates of the redshift z_{MAP} are obtained by taking the maximum *a posteriori* (MAP) value of each constituent redshift posterior, marginalising over the other three parameters. The method recovers the true redshift of each constituent from simulated photometry well. The performance of

photometric redshift methods is often summarised by the RMS scatter $\sigma_{\rm RMS}$. We first define the normalised error for galaxy g as

$$\tilde{\delta} z_g = \frac{\hat{z}_{s,g} - z_{{\rm MAP},g}}{1 + \hat{z}_{s,g}} \,, \tag{6.44}$$

where each galaxy g is a single constituent of a blended source. Writing the total number of galaxies in our test catalogue as N_g, we then define the RMS scatter as

$$\sigma_{\rm RMS} = \sqrt{\frac{1}{N_g} \sum_g \left(\tilde{\delta} z_g \right)^2} \,. \tag{6.45}$$

Computing this quantity for our mock blended observations, we find an RMS scatter of $\sigma_{\rm RMS} = 0.163$. This compares to a scatter of $\sigma_{\rm RMS} = 0.0267$ when testing the method with $N = 1$ constituents on mock observations of single sources. In this case, the method almost reduces to the BPZ (Benítez, 2000) formalism, the only difference being the magnitude prior which has a negligible effect in the $N = 1$ case. We verified that this was the case by performing a test on the same $N = 1$ simulated data using the BPZ code, finding a negligible difference between the results from BPZ and `blendz`.

This scatter can be improved by excluding sources with photometric redshifts that, using the uncertainty information of the posterior distribution, are identified as untrustworthy. This is done by comparing a summary statistic against a threshold that controls the stringency of the test; we use the standard deviation of redshift marginal-posterior samples σ_α separately for each constituent, though a variety of summary statistics are available. Keeping only sources with $\sigma_\alpha \leq 0.2 \, \forall \, \alpha$, the RMS scatter is reduced to $\sigma_{\rm RMS} = 0.064$, with 37% of sources removed. The effect of this is shown in the bottom row of Figure 6.9.

The percentage of outliers can also be quantified. Outliers are defined as sources where either constituent has an error $|z_{\rm MAP} - \hat{z}_s| \geq 0.15(1 + \hat{z}_s)$. For the full set of mock observations, this percentage was found to be 18.6%. By keeping only sources with $\sigma_\alpha \leq 0.2 \, \forall \, \alpha$ as described above, the percentage of outliers falls to only 6.0%.

The results of the detection of blends are also shown in the centre panels of Figure 6.9. By using equation 6.43, we calculate $\mathcal{P}_{2,1}$, the relative probability that a source is a two-constituent blend compared to a single source. The interpretation of this probability is problem-dependent; a probability of $\ln(\mathcal{P}_{2,1}) > 0$ indicates a preference towards the source being blended, while a threshold to $\ln(\mathcal{P}_{2,1}) > 5$ indicates strong evidence (Kass and Raftery, 1995). Likewise, probabilities of $\ln(\mathcal{P}_{2,1}) < 0$ and $\ln(\mathcal{P}_{2,1}) < -5$ indicate a preference and strong evidence for the single source

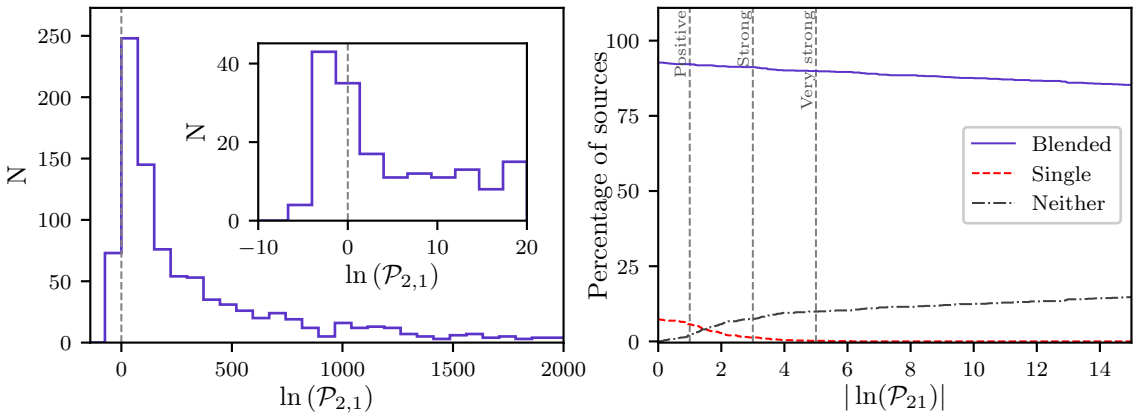

Figure 6.10: The left panel shows the distribution of the relative blend-to-single probability for the mock catalogue, with the inset showing the same distribution, zoomed around lower relative probabilities and binned more finely. The right panel shows the percentage of sources assigned as either blended, single sources or not assigned to either as the threshold for deciding between each label is changed.

case respectively. As the blended and single thresholds are pushed more positive and negative respectively, there are sources with values of $\ln(\mathcal{P}_{2,1})$ that fall between these thresholds. In these cases, the source is assigned neither label.

As described in section 6.3.1, we assume the relative prior probability of a blend to be $P(N = 2)/P(N = 1) = 1$, i.e., we give no preference to either model. Under this assumption, the method identifies 92.7% of sources as blends and 7.3% as single sources. Increasing the threshold to strong evidence, we find that 89.9% of sources are identified as blends and 0.2% as single sources; the remaining 9.9% fall between these thresholds. The distribution of the relative probability of blending and the effect on blend identification of changing the threshold are shown in Figure 6.10. We also performed the same test for $N = 1$ constituent simulated data, and found that the method identified 96.3% of these sources as not being blended.

These results show that the method can both recover the redshifts from broad-band observations of blended objects, and detect the blending of a large fraction of these objects from their photometry alone. In addition, the output from these tests are not just point estimates of redshifts, but the full four-dimensional posterior distributions that capture the correlations between constituents that can be lost by working with constituent separated maps.

These are the results of simulated observations, however; real data has the complication that the flux model is no longer exact, i.e., the templates are not perfectly representative of all galaxies observed. As such, we test the method on real data in section 6.5.

6.4.2 Partially-blended sources

To test the effect of adding resolved data, we created a set of mock photometric observations of two-constituent partially blended systems. These observations simulate the same six-band optical survey as described in section 6.4.1, combined with a four-band optical and infrared space-based survey using the Euclid filters vis, Y, J, H (Racca et al., 2016). This latter survey is assumed to have made resolved measurements of each constituent, while the former is fully-blended as before. Thus, our partially-blended data vector contains 14 fluxes for each source.

For comparison with the partially-blended results, we repeat the inference several times. Firstly, we compare against the fully-blended LSST-like case described above. Next, we compare to an inference using the resolved Euclid bands only, testing the effect of removing the difficulty of blending but using lower signal-to-noise data. Finally, we test against the case of using both the LSST- and Euclid-like data, but assuming that sources are blended in all bands. This allows us to separate the improvement as a result of adding resolved data from that of simply having more bands available.

For the fully-blended bands, we reuse the simulated fluxes described in section 6.4.1. For the resolved bands, we generate observed fluxes using the same randomly sampled source parameters. The observed fluxes are then generated using the flux model described in section 6.1.1 with added observational errors drawn randomly from an uncorrelated, zero centred normal distribution. The noise in these resolved bands is set to the final 1σ depths expected from Euclid observations (Laureijs et al., 2011).

We use the same prior as described in section 6.4.1 for both the simulation and inference steps. The reference band over which the prior is defined is set to be the r-band of the blended observations. This band is not present during the inference step using only the resolved data. As a result, we use the flux model from section 6.1.1 to convert between r- and Y-band magnitudes before evaluating the prior.

Figure 6.11 shows a comparison of the photometric redshift point estimates with the true simulated values for the four sets of inferences. As before, these point estimates are the MAP values of the marginal redshift posteriors.

The left panel shows the fully blended case, the same results as section 6.4.1. Again, the redshifts of many constituents are well recovered, though there is a significant fraction of outliers. The RMS scatter in this case was found to be $\sigma_{\mathrm{RMS}} = 0.163$. The percentage of outliers, sources where either constituent has an error $|z_{\mathrm{MAP}} - \hat{z}_s| \geq 0.15(1 + \hat{z}_s)$, was found to be 18.6%.

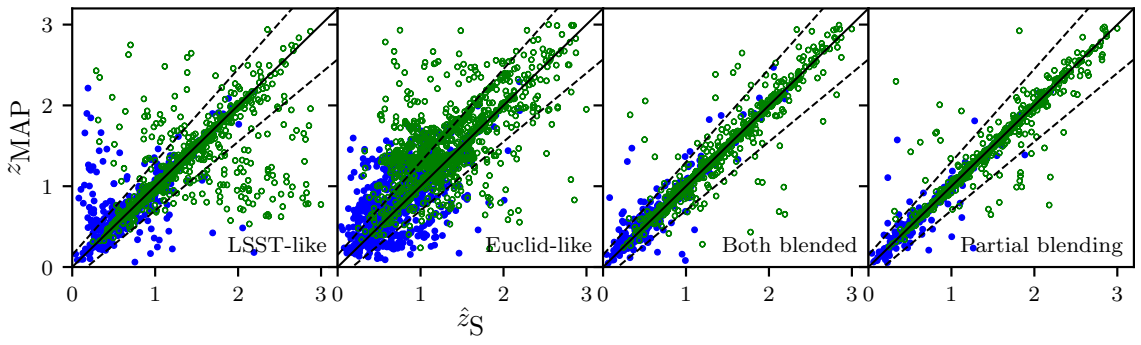

Figure 6.11: Scatter plot comparing the maximum *a posteriori* point estimates for the fully-blended, resolved and partially-blended cases. The closed blue markers represent the redshift of the closer constituent, z_α, while the open green markers represent the redshift of the more distant constituent, z_β.

The centre-left panel of Figure 6.11 shows the results for the resolved observations. Though finding the photometric redshift of resolved sources is an easier inference problem, this is counteracted by the significant reduction in the signal-to-noise of this data. As a result, we find an RMS scatter of $\sigma_{\mathrm{RMS}} = 0.212$ with 55.0% of sources marked as outliers.

The centre-right panel of Figure 6.11 shows the results for combination of LSST- and Euclid-like data in the fully blended case. We find that the addition of the four Euclid bands significantly improves the precision of the redshift inference, which has an RMS scatter of $\sigma_{\mathrm{RMS}} = 0.073$. The fraction of outliers has also improved significantly, with only 6.6% of sources marked as outliers.

Finally, the right panel of Figure 6.11 shows the results for the partially blended case, combining the high-precision blended observations with the resolved data. Here, we find that the RMS scatter has reduced to $\sigma_{\mathrm{RMS}} = 0.065$, a factor of 2.5 improvement over the blended LSST-like data alone, and a factor of 1.12 over the combined LSST- and Euclid-like blended data. The percentage of outliers has also been reduced. Here, only 3.4% of sources are found to be outliers, a factor of 5 improvement over the fully-blended case, and a factor of 1.9 improvement over the combined LSST- and Euclid-like blended data.

While the most significant improvement was obtained through the increase in the number of bands, these results show that the quality of photometric redshifts of blended sources can be improved through the inclusion of resolved data. This is particularly apparent in the reduction of outliers. One advantage conferred by the addition of resolved data is a constraint on the relative magnitudes of blended constituents. In the fully-blended case, the reference-band magnitude of each of the constituents must be

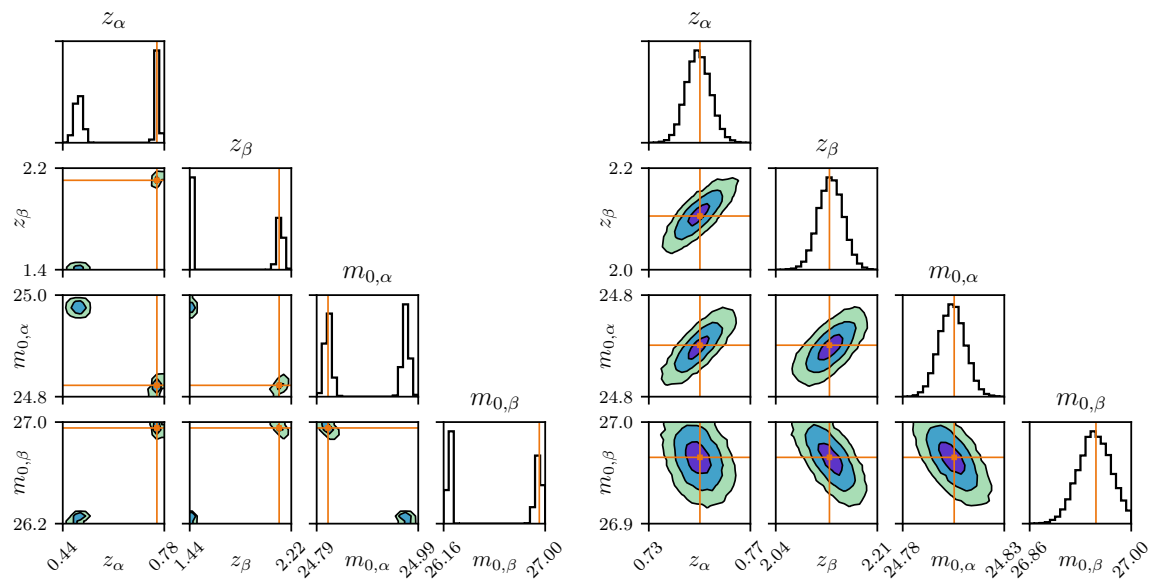

Figure 6.12: The 4D posterior distributions for a two-constituent blended source in the fully-blended and partially-blended cases. The left plot shows the result of inference using blended data only. While their is significant posterior density around the true parameter values shown by the orange line, this posterior is highly bimodal, with two distinct solutions that cannot be distinguished. The plot on the right shows the result of the partially-blended case that includes both blended and resolved observations. The addition of information about the magnitude of each constituent separately has removed the incorrect mode, resulting in a posterior that recovers the true solution well.

inferred from the combined magnitude of the blended source only. This can lead to the degenerate distributions shown in Figure 6.8. Adding resolved photometry can help to break this degeneracy by providing information about each constituent individually. The precision of the photometric redshift inferences is therefore improved.

An example of this phenomenon is shown in Figure 6.12. The left panel shows a corner plot of the posterior distribution for a fully-blended source. The marginal distributions for each constituent redshift are highly multimodal, with well separated redshifts occurring at distinct magnitudes. Though there is significant posterior density around the true redshifts, the MAP point estimate of z_β would show a significant error, as the incorrect mode has a higher posterior. The right panel of Figure 6.12 shows the same source analysed in the partially-blended case after the addition of the resolved photometry. Here, the width of the posterior has been significantly reduced by the removal of the incorrect mode. The posterior now shows that the redshift of the source has been well constrained, and the redshift point estimates would no longer have a large error.

6.5 GAMA blended sources catalogue

The Galaxy And Mass Assembly (GAMA) survey (Baldry et al., 2017) is a spectroscopic galaxy survey that observed 286 deg^2 of sky over several regions to a magnitude limit of between $r < 19$ and $r < 19.8$. In doing so, it obtained precise redshifts of $> 150\,000$ sources. The observed regions were chosen to overlap with existing imaging surveys such as Sloan Digital Sky Survey (SDSS) (Stoughton et al., 2002) and VISTA Kilo-degree Infrared Galaxy (VIKING) Survey (Edge et al., 2013). As a result, the spectroscopic data is accompanied by a set of aperture-matched photometry covering nine filter bands $u, g, r, i, z, Y, J, H, K$ from optical to infrared wavelengths (Hill et al., 2011).

The GAMA blended sources catalogue (Holwerda et al., 2015) contains 280 sources from the GAMA survey that have been spectroscopically identified as blended objects. These were selected using an automated template-based spectrum fitting method (Baldry et al., 2014) that cross correlates galaxy templates with the observed spectra to determine the galaxy redshift. Sources where two different redshifts showed strong cross-correlations were visually inspected, resulting in a selection of blended galaxies. The motivation of Holwerda et al. (2015) was the identification of strong lens candidates. However, a catalogue of spectroscopically identified blended galaxies with accompanying nine-band photometry gives us an useful test case for the blended photometric redshift estimation method on non-simulated photometry with secure redshifts available for both constituents.

We first calibrate the prior using the procedure described in section 6.1.6. To do this, we used 26782 unblended, well-observed galaxies. These were selected by enforcing every band to be free from SExtractor (Bertin and Arnouts, 1996) error flags and excluding all galaxies in the blended source catalogue. The resulting prior from the calibration procedure is shown in Figure 6.6. As discussed in section 6.1.5, we test two methods of setting the faint-end magnitude cut $m_{\max}$, firstly as a $5\sigma_0$ flux deviation using equation 6.35, and secondly, fixing $m_{\max} = 20.8$. Throughout, we refer to these as the sigma-$m_{\max}$ case and fixed-$m_{\max}$ case respectively. We then proceed with the inference using the same template set[3] described in section 6.4.

The resulting redshift point estimates are shown in Figure 6.13. While noisier than the simulated case, the method still recovers reasonable estimates; using equation 6.45, we find an RMS scatter of $\sigma_{\mathrm{RMS}} = 0.156$ in the both the sigma-$m_{\max}$ and

[3]The templates used to fit the spectroscopic redshifts as described in Holwerda et al. (2015) do not cover the full wavelength range of the photometry. As a result, we do not use them for the photometric redshift inference.

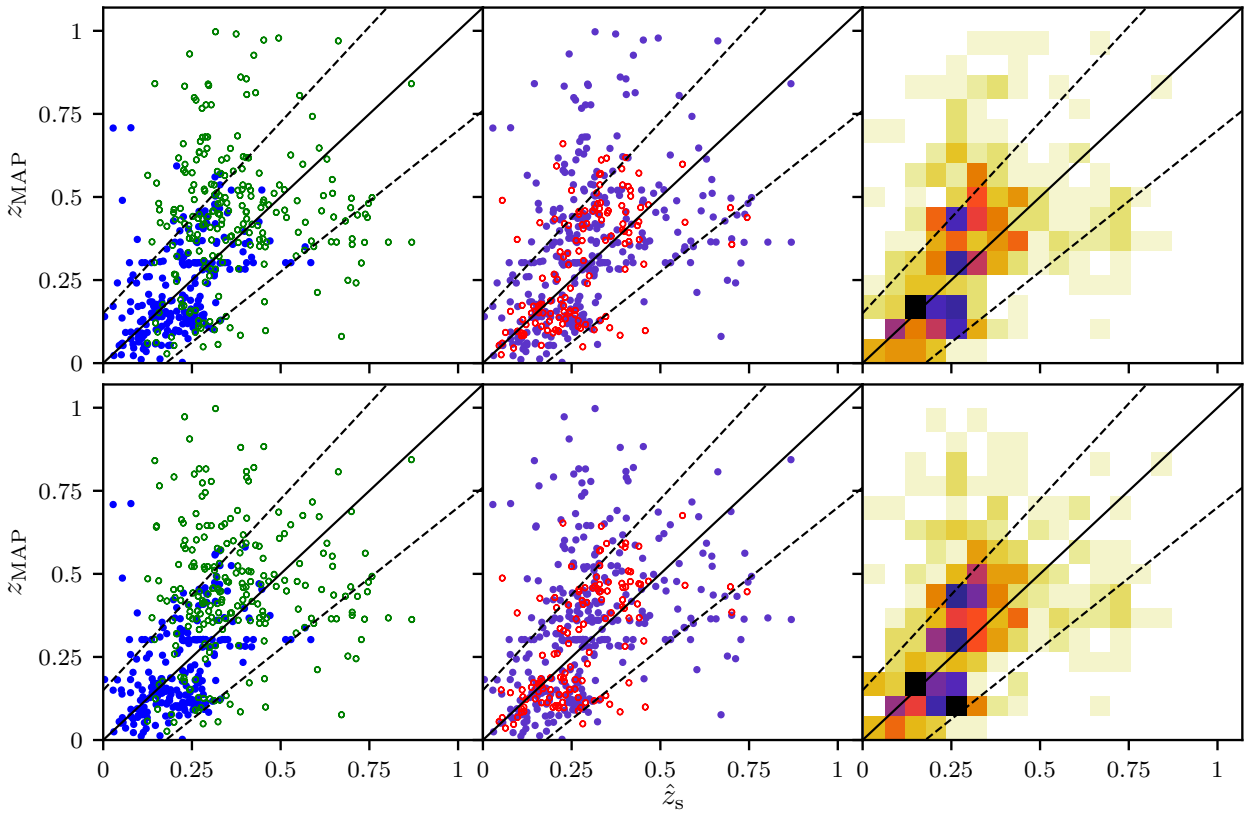

Figure 6.13: Scatter plot comparing the maximum *a posteriori* point estimates from the photometric redshift estimation with the spectroscopic redshifts for sources from the GAMA blended sources catalogue. The left panels distinguish the constituents, with z_α plotted with closed blue markers, and z_β plotted with open green markers. The centre panels show the blend identification, with sources identified as blends plotted with closed purple markers, and those misidentified as single sources plotted with open red markers. The right panels show a 2D histogram of the combined sample. Panels in the top row show the results for the sigma-m_{max} case, while those in the bottom row show the fixed-m_{max} case. The dashed lines in each panel show an error of $0.15(1+z)$.

fixed-m_{max} cases. This compares to the scatter for a set of unblended GAMA sources of $\sigma_{\mathrm{RMS}} = 0.116$. Obtaining this value even without the added complication of blending suggests a mismatch between the sources and the template set.

We also compute the inferred blend probability $\mathcal{P}_{2,1}$ for these galaxies. The distribution of these probabilities is shown in Figure 6.14. As described in section 6.4, $\ln(\mathcal{P}_{2,1}) > 0$ and $\ln(\mathcal{P}_{2,1}) > 5$ show a preference and strong evidence for a blended source respectively, while $\ln(\mathcal{P}_{2,1}) < 0$ and $\ln(\mathcal{P}_{2,1}) < -5$ show the same for the single source case. The distribution of the blend probability and the effect of the evidence threshold on blend identification is shown in Figure 6.14.

In our tests of the sigma-m_{max} case, 71.6% of sources showed a preference for being blended, with 28.4% preferring a single source. Increasing the threshold to

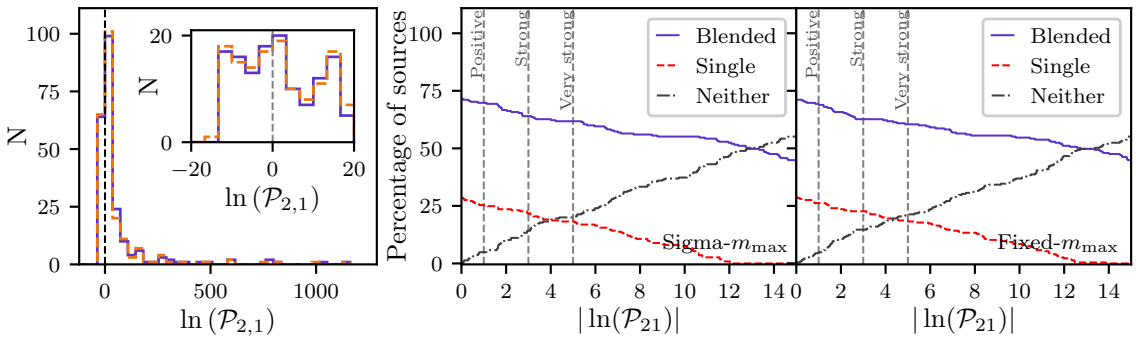

Figure 6.14: Plots showing the differences in the model comparison results between the two methods tested of setting the faint-end magnitude cut $m_{\rm max}$, labelled the sigma-$m_{\rm max}$ and fixed-$m_{\rm max}$ cases. The left panel shows the distribution of the relative blend-to-single probability for the mock catalogue, with the inset showing the same distribution, zoomed around lower relative probabilities and binned more finely. The solid line shows the sigma-$m_{\rm max}$ case, and the dashed line shows the fixed-$m_{\rm max}$ case. The two right panels shows the percentage of sources assigned as either blended, single sources or not assigned to either as the threshold for deciding between each label is changed.

strong evidence, these percentages fall to 61.8% and 18.2% respectively. Finally, the incorrectly identified single sources can be excluded entirely by increasing the threshold to $|\ln(\mathcal{P}_{2,1})| < 12.5$, with 50.7% of sources identified as blends at this level.

The identification of blends was very similar in the fixed-$m_{\rm max}$ case. We found that 71.1% of sources showed a preference for being blended, and 28.9% preferred a single source. At the strong evidence threshold, 60.4% of sources are correctly identified as blends, with 18.2% misidentified as single sources. The threshold to exclude misidentified sources completely in the fixed-$m_{\rm max}$ case is $|\ln(\mathcal{P}_{2,1})| < 13.9$, slightly higher than the sigma-$m_{\rm max}$ case. At this level, 48.0% of sources are still correctly identified as blends.

These results show that photometric redshift estimates can be obtained for blended sources, and that the method can identify many blended sources from just their broad-band photometry. By adjusting the threshold of the probability $\mathcal{P}_{2,1}$, blended sources can be selected in a way that trades off completeness and purity.

Several techniques for improving the scatter of photometric redshifts have been proposed, such as rest-frame template error functions (Brammer et al., 2008), iterative methods to modify templates to be more representative (Feldmann et al., 2006), using clustering-based redshift estimation to calibrate systematic biases using galaxies (Gatti et al., 2018) and intensity mapping observations (Alonso et al., 2017), and constructing priors in terms of physical galaxy properties (Tanaka, 2015). While an investigation of

these methods is beyond the scope of this chapter, they could also be applied while using this method. This could help to reduce the scatter of the blended photometric redshift estimates to a level necessary for future surveys, while retaining the full information of the posterior for accurate error propagation.

6.6 Conclusions

Blended sources will become far more common in future galaxy surveys than are found currently due to increases in the depth of photometry and as a result, the number density of galaxies. This chapter presents a Bayesian photometric redshift method that generalises the existing BPZ (Benítez, 2000) method to the case of blended observations. We derive a posterior for the redshift and magnitude of each constituent which we sample to obtain estimates of the redshift. We also use this posterior in a model comparison procedure to infer the number of constituents in a source.

By doing this, the method is able to infer both the redshift of each constituent within a blended source, and identify that a source is blended from its broadband photometry alone. The joint posterior distribution of the redshifts of all constituents in a blend provides a complete accounting of the correlations in the final result, information that can be lost when separating constituents and estimating redshifts for each separately. This uncertainty information is essential for obtaining accurate uncertainties on cosmological parameters that rely on the photometric redshift estimates. A Python implementation of the method, `blendz`, is available to download.

By inferring the redshifts of constituents directly from their blended photometry, the method presented here is directly applicable to ambiguously blended objects that cannot otherwise be deblended. The partial-blending formalism described in section 6.2 also enables the catalogue-level joint analysis of sources in space- and ground-based surveys such as Euclid and LSST. The complementarity of these surveys will allow cosmological parameters to be constrained more precisely than either survey could individually, and analysis of blended sources from their aperture photometry will be simpler than a joint pixel-level analysis (Rhodes et al., 2017).

The method presented here could also be combined with existing deblending methods that utilise the spatial information of images directly. These methods are complementary; image-based deblending methods are effective provided that constituents are sufficiently well separated. If this is not the case, there is too little spatial information to be able to separate constituents, and colour information is necessary. Combining these methods could allow future surveys to identify a greater proportion of blended

sources, reducing their effects on cosmological constraints. Deblending methods that also incorporate colour information would need to be combined with this method more carefully however, as the colour information would be used twice and thus the blending probabilities would not be independent. This method could instead be extended to incorporate imaging data by constructing a forward model of the galaxy in each band and constraining both morphology and redshift simultaneously. This is discussed further in chapter 9.

The method presented here is focussed on the problem of galaxy-galaxy blending. However this formalism could also be applied to the problem of inferring photometric redshifts of galaxies blended with other objects such as stars and quasars with minimal modifications. For the latter problem, only the inclusion of quasar templates would be required. If these templates were included alongside the existing galaxy templates as a single template set, the marginalisation over templates would then also implicitly be a marginalisation over the classification of each object into either a galaxy or a quasar. The prior probability of the classification of each constituent could then be absorbed into the template prior.

If a marginalisation over object classification was not desired, one could specify a choice of object type for each constituent and marginalise over only the relevant template sets for each. This would effectively make the template prior object-classification dependent, e.g., the template prior for a galaxy template given a quasar classification would equal zero. In this case, the object classification of each constituent would become another model that could be compared with other choices using the same model comparison procedure described above.

A similar procedure could be applied for star-galaxy blending. However, in this case, the redshift of stars can be assumed to be negligible compared to that of galaxies. As a result, the redshift of the relevant constituent would no longer appear as a parameter in the model, and the star would be described by only its magnitude and template.

Chapter 7

Gaussian Mixture Models for Blended Photometric Redshifts

This chapter is heavily based on work from Jones and Heavens (2019b).

An alternative to deblending is to infer quantities of interest, such as photometric redshifts, from blended data directly. This joint approach automatically accounts for correlations between each galaxy in a blended source and correctly propagates these uncertainties to the final results. This is the approach taken in chapter 6, which generalises Bayesian template-based photometric redshift methods to the case of blended observations.

In order to sample the posterior and evaluate the evidence, chapter 6 uses MultiNest (Feroz et al., 2009), an efficient implementation of the nested sampling method (Skilling, 2006). However, even sampling with an efficient method such as MultiNest can be computationally demanding; sampling the posteriors for both one- and two-constituent models takes approximately two minutes per source on a workstation with a 3 GHz Intel Xeon processor. While this is viable for small samples, it is not scalable to the large samples of $\sim 10^9$ galaxies in a future survey like LSST.

This chapter takes the same joint-inference approach, but uses a Gaussian mixture model to learn the flux-redshift relation from a training set of galaxies with known redshifts. We then use this model as a prior to derive the posteriors and marginal-likelihoods for sources consisting of one or two galaxies. Since these can be computed analytically, this is significantly less computationally demanding than the nested sampling-based method described in chapter 6, an important property for use in future galaxy surveys. As a result, the one- and two-constituent inference and model selection can be done for approximately ten sources per second, a speed-up of three orders of

Table 7.1: A summary of the notation used throughout this chapter.

Symbol	Description
N	Number of constituent galaxies in a source
z_n	Model redshift of constituent galaxy n
$\boldsymbol{F}_n$	Model flux vector of constituent galaxy n
$\hat{\boldsymbol{F}}$	Vector of observed fluxes
$\underline{\boldsymbol{\Sigma}}^{\hat{F}}$	Covariance matrix of observed fluxes
M	Number of components in the mixture model
w^k	Weight of mixture component k
$\boldsymbol{\mu}^k$	Mean vector of mixture component k
$\underline{\boldsymbol{\Sigma}}^k$	Covariance matrix of mixture component k
$\mathcal{E}^1$	Evidence for single-constituent model
$\mathcal{E}^2$	Evidence for two-constituent model
$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \underline{\boldsymbol{\Sigma}})$	Multivariate Gaussian PDF with mean vector $\boldsymbol{\mu}$ and covariance matrix $\underline{\boldsymbol{\Sigma}}$
$\tilde{\mathcal{N}}(\boldsymbol{x}\|\boldsymbol{\eta}, \underline{\boldsymbol{\Lambda}})$	Multivariate Gaussian PDF in natural parametrisation with parameters $\underline{\boldsymbol{\Lambda}} \equiv \underline{\boldsymbol{\Sigma}}^{-1}$ and $\boldsymbol{\eta} \equiv \underline{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}$

magnitude on the workstation described above. Photometric redshift inference is also trivially parallelisable for high-performance computing environments, since each source can be considered independently.

Throughout this chapter, we use the term *constituent* to describe the individual galaxies comprising a blended source. Following convention, we refer to each multivariate Gaussian distribution in the mixture model as a *component*. We denote scalars using an italic font x, vectors using a bold italic font $\boldsymbol{x}$ and matrices using a bold sans-serif font $\underline{\boldsymbol{x}}$. We summarise our notation in Table 7.1.

This chapter is organised as follows. In section 7.1, we introduce our formalism for blended photometric redshifts with Gaussian mixture models. We use this to derive expressions for the posteriors and evidences in section 7.2. We present results of tests of our method on simulated data in section 7.3. Finally, in section 7.4, we present these tests on real blended data from the Galaxy And Mass Assembly (GAMA) survey (Baldry et al., 2017).

7.1 Gaussian mixture model photo-z

Photometric redshifts inferred using machine learning methods are often very accurate when good training data is available. These methods perform regression, and use this training data to learn the mapping from fluxes to redshifts. Many machine learning algorithms are not inherently probabilistic; a particular input will map to a particular output. However, accurate uncertainties on cosmological parameters rely on propagating uncertainties from all stages of the analysis. Machine learning photometric redshift methods have therefore developed several ways to estimate these uncertainties.

One example that accounts for errors in the observed fluxes is to apply the chain rule to successive layers of a neural network (Collister and Lahav, 2004), providing the variance of the output redshift. Some machine learning methods such as a Gaussian process (e.g., Way and Srivastava, 2006), are already explicitly probabilistic, naturally producing variance estimates alongside their prediction. Other methods can represent their uncertainties more generally by inferring PDFs as their output. This can be done by training many machine learning algorithms to each independently estimate the redshift and taking the distribution of the ensemble to be the redshift PDF (Sadeh et al., 2016). A single neural network can also accomplish this by being trained to output the parameters of a parametrised PDF rather than the redshift directly (D'Isanto and Polsterer, 2018). PDFs represent the complete probabilistic knowledge over a system under investigation, and are thus a general mechanism for quantifying and propagating uncertainties within a statistical analysis (e.g., Gelman et al., 2013).

In addition to enabling the rigorous propagation of uncertainties, using full photometric redshift PDFs has been shown to improve the accuracy of cosmological analyses (e.g., Mandelbaum et al., 2008; Myers et al., 2009). PDFs also have an advantage over simply representing uncertainty with the variance in their ability to represent multimodality; that is, several distinct, well separated redshifts being plausible for a given vector of fluxes. This is a common occurrence in photometric redshifts (Benítez et al., 2009). Colour-redshift degeneracies mean that high- and low- redshift galaxies can have similar colours, often due to spectral features such as the Lyman and Balmer breaks being misidentified as one another (Graham et al., 2018).

Here, we treat the training data not as variables to regress between, but instead as noisy samples from the joint redshift-flux distribution, turning the problem into one of density estimation. The joint density is the most general probabilistic description of the training data, allowing several quantities of interest to be derived. Given an observed vector of fluxes $\hat{\boldsymbol{F}}$, the redshift can be inferred using the conditional distribution $P(z \mid \hat{\boldsymbol{F}})$ which can be derived from the joint distribution. This PDF can be

multimodal, capturing the degeneracy described above. These distributions can be composed together to produce the conditional distribution of the redshifts of a blended source $P(z_1, z_2 \mid \hat{\boldsymbol{F}})$ in a similar fashion. The joint distribution also permits calculation of marginal likelihoods, allowing Bayesian model selection techniques to be used to infer the number of constituents in a source. Finally, the interpretation of the joint distribution is clear, in contrast to other machine learning methods that can be 'black-boxes', requiring additional ad-hoc techniques to improve their interpretability (e.g., Shrikumar et al., 2017; Shwartz-Ziv and Tishby, 2017)

We model the joint distribution of the latent, noise-free parameters as a Gaussian mixture model (GMM), a weighted linear combination of multivariate Gaussians, i.e.,

$$P(z, \boldsymbol{F}) = \sum_k w^k \, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) \, . \tag{7.1}$$

By imposing that $\sum_k w^k = 1$, this density is correctly normalised, i.e.,

$$\sum_k w^k \iint \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) \, \mathrm{d}z \, \mathrm{d}\boldsymbol{F} = \sum_k w^k = 1 \, . \tag{7.2}$$

This choice has several useful features. Firstly, GMMs are easy to train using standard, well-tested methods. This is discussed further in section 7.1.1. Secondly, inference with GMMs is computationally inexpensive as they can be efficiently sampled as detailed in section 7.1.4. Lastly, GMMs are mathematically convenient. Both the conditional and marginal distributions of multivariate Gaussians are also Gaussians. The same is also true of both the product and convolution of several multivariate Gaussians. These properties will be used frequently throughout this chapter to render many calculations analytic. Despite this, GMMs can represent a wide variety of PDFs, including those that are skewed or multimodal. This is demonstrated in Figure 7.1.

Using GMMs to infer photometric redshifts in this way was first done in Bovy et al. (2012), who applied the method to obtain photometric redshifts of quasars and used model selection techniques to separate stars and quasars. The method we present in this chapter differs from this in several ways. Firstly, we extend the method to the case of jointly inferring multiple redshifts directly from blended data.

Secondly, Bovy et al. (2012) fit a series of many GMMs to the fluxes and redshifts of quasars in several magnitude bins. As a result, our model has significantly fewer parameters to fit. Nevertheless, our use of cross-validation to set the number of mixture components as described in section 7.1.3 provides the model sufficient flexibility to fit the flux-redshift density with the full fidelity provided by the training set.

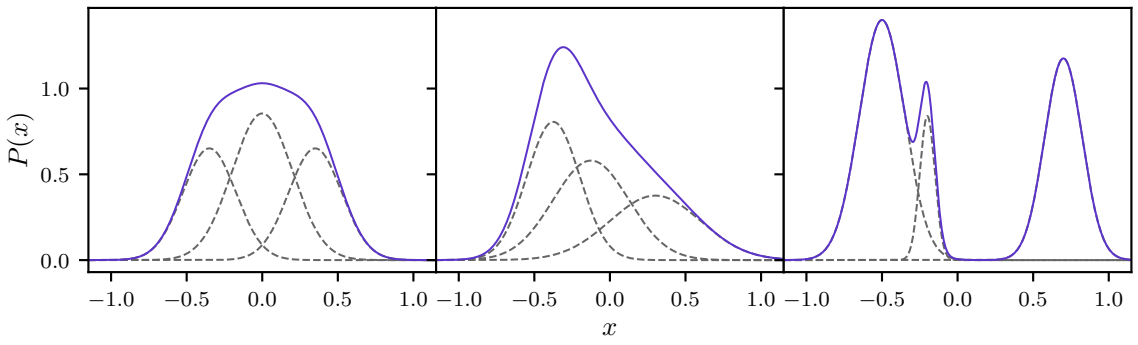

Figure 7.1: Plot showing a variety of PDFs that can be represented by Gaussian mixture models, given a sufficient number of components. The dashed grey curves show each weighted Gaussian component, and the solid blue curves show the mixture formed by the linear combination of these components.

The binning of Bovy et al. (2012) is not possible due to the extension to blended sources. Observations in this case are of the flux of the blended source, while the magnitude bin in that model is chosen based on the magnitude of an individual galaxy. This quantity that is not observed in the blended case, and so cannot be used to choose a magnitude bin. The same is true of colours, i.e., ratios of fluxes relative to the flux in a particular reference band, which are often used in machine learning-based photometric redshift methods. Since the reference-band flux of each galaxy in a blended source is not observed, the colours for each galaxy cannot be calculated and so cannot be used to infer the redshifts.

Finally, our derivation does not use the convolution property of multivariate Gaussians described above, since integrals over fluxes are then implicitly evaluated from $-\infty$ to ∞ as multivariate Gaussians have infinite support. These integrals therefore contain contributions from non-physical negative fluxes. This is a safe approximation when considering unblended sources, since their flux is strongly constrained by observations. However, the same is not true of blended sources, where the individual flux of each constituent is not observed. Instead, we evaluate these results using an efficient Monte Carlo integration method. We therefore treat unblended sources in the same way for consistency.

All fluxes throughout are renormalised for numerical stability. This is done by dividing each flux by the standard deviation in the training set, e.g., for band b,

$$F_b \to \frac{F_b}{\sigma(\{\hat{\boldsymbol{F}}_b\}_{\mathrm{tr}})} \,. \tag{7.3}$$

Normalising the data in this way is a common preprocessing step in machine learning

methods. Without this renormalisation, the observed fluxes are small enough that the EM fitting procedure is dominated by numerical errors as the covariance matrices of the components become poorly conditioned. The corresponding change in the covariance matrix of each data point is given by

$$\Sigma_{ij} \to \frac{\Sigma_{ij}}{\sigma(\{\hat{\boldsymbol{F}}_i\}_{\mathrm{tr}})\sigma(\{\hat{\boldsymbol{F}}_j\}_{\mathrm{tr}})} \,. \tag{7.4}$$

We also note that magnitudes are commonly used for this purpose in machine learning-based photometric redshift methods, since the logarithmic transformation of the flux also effectively normalises them. However, an advantage of the GMM method presented here is that expressions for posteriors and evidences can be calculated analytically. This relies on the model for the flux of the blended sources being a linear combination of the fluxes of the individual constituents, since this leaves the likelihood of the sum a Gaussian. This would no longer be the case when using magnitudes, as the model for the magnitude of a blended source would be a non-linear function of the individual constituent magnitudes.

7.1.1 Training Gaussian mixture models

Our prior density $P(z, \boldsymbol{F})$ is defined in terms of the true, latent parameters. Therefore, this density must be fitted with a method that incorporates both the noisy data and the covariance. To do this, we use extreme deconvolution (Bovy et al., 2011), an extension of the expectation-maximisation (EM) algorithm (Dempster et al., 1977) commonly used the find the maximum-likelihood parameters of GMMs. This is the same fitting method as the quasar photometric redshift method of Bovy et al. (2012).

Extreme deconvolution generalises the EM algorithm to the case where the data is subject to normally-distributed errors. The EM algorithm is a general method for fitting models with some form of hidden data in addition to the observed data. Given an initial guess at the parameters, the algorithm iteratively modifies these parameters to increase the likelihood, converging to a local maximum.

For a single multivariate Gaussian, the maximum-likelihood parameters can be found exactly through the derivative of the likelihood. However, the same is not true of mixtures of Gaussians, as these parameters are not available in closed form. The hidden information that would make this tractable is the identity of the component from which each sample was drawn. If this were known, fitting the GMM would reduce to the previous analytic case. Though this information is hidden, this points

to an iterative solution; first, the parameter guess can be used to update the hidden information, then this information can be used to update the parameters.

In essence, expectation-maximisation is a probabilistic version of this procedure that takes into account the uncertainty in the hidden information. By averaging the likelihood over the probability of each sample being drawn from each component, the maximum likelihood parameters can be found in closed form. Since the component probability depends on the parameters being fitted, this process is iterative.

The extreme deconvolution method of Bovy et al. (2011) extends the EM algorithm to fit data with Gaussian errors. This is done by replacing the likelihood with a marginalised version given by

$$P(\hat{\boldsymbol{x}} \mid \{\theta\}) = \int P(\hat{\boldsymbol{x}}, \boldsymbol{x} \mid \{\theta\}) \, \mathrm{d}\boldsymbol{x} = \int P(\hat{\boldsymbol{x}} \mid \boldsymbol{x}) P(\boldsymbol{x} \mid \{\theta\}) \, \mathrm{d}\boldsymbol{x} \,, \tag{7.5}$$

where $\hat{\boldsymbol{x}}$ is the vector of observed values, $\boldsymbol{x}$ is the latent vector of true values and $\{\theta\}$ are the mixture parameters being fitted, i.e., weights, means and covariances. The data likelihood $P(\hat{\boldsymbol{x}} \mid \boldsymbol{x})$ is assumed to be a multivariate Gaussian, and $P(\boldsymbol{x} \mid \{\theta\})$ is the GMM. Due to the convolution property of multivariate Gaussians, this marginalised likelihood is also a Gaussian mixture, and thus amenable to being fitted using an expectation-maximisation approach. Using this extreme deconvolution method, we fit the joint flux-redshift distribution $P(z, \boldsymbol{F})$ while accounting for uncertainties in the training set.

This fitting procedure assumes that the number of mixture components is fixed. The method we use to decide on this number is discussed in section 7.1.3.

As discussed above, multivariate Gaussians have infinite support, and so non-physical negative fluxes and negative redshifts are *a priori* allowed. No non-physical fluxes will be present in the training set, and negative redshifts, while not non-physical, are sufficiently rare that they can be presumed to not be present either. As a result, there is no incentive for the training algorithm to assign significant prior volume here. However, without an additional prior on the mixture parameters, prior volume in negative regions is not penalised either.

It is possible to generalise EM-based methods such as extreme deconvolution to maximise the posterior rather than the likelihood by adding a log-prior. However, while this will ameliorate the problem of negative values, it cannot eliminate it completely; the GMM having infinite support means that every point in parameter space will always have non-zero density.

An alternative approach is to impose an additional prior that is zero is any

negative regions of parameter space, i.e.,

$$P(z, \boldsymbol{F}) = \psi(z, \boldsymbol{F}) \sum_k w^k \, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) \tag{7.6}$$

where

$$\psi(z, \boldsymbol{F}) = \begin{cases} 0 & \text{for } z, \boldsymbol{F} < 0 \\ 1 & \text{otherwise.} \end{cases} \tag{7.7}$$

This will exactly fix the problem of negative values. However, it will also force otherwise analytic integrations to have to be done numerically. These cases are discussed in the relevant sections below.

Imposing this boundary prior will also change the normalisation of the prior from unity, i.e.,

$$\iint \psi(z, \boldsymbol{F}) \sum_k w^k \, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) \, \mathrm{d}z \, \mathrm{d}\boldsymbol{F} \neq 1 \, . \tag{7.8}$$

The model selection described below requires that the prior be normalised. This normalisation differs between the single- and two-constituent cases, with the latter also being affected by the sorting condition. These normalisations are therefore discussed in their respective sections below.

It should be noted that, since this is an empirical method that does not rely on any underlying physical model in the way that a template-based method does, the redshift can be transformed almost arbitrarily. The only restrictions in this transformation are that it is both invertible and well-defined for all positive real numbers. The only modifications to the method required to accommodate this are to the limits of redshift integrals. For a transformation $\mathcal{T}(z)$, the lower and upper limits should be replaced with $\mathcal{T}(0)$ and $\mathcal{T}(\infty)$ respectively.

The transformation $\mathcal{T}(z) = \log(z)$ would seem to be a sensible choice, as the lower and upper integration limits would become $-\infty$ and ∞ respectively, rendering all the redshift integrations throughout analytic. This is the approach taken by Bovy et al. (2012). However, in our tests, we found that this transformation reduces the accuracy of the blended redshift inference. The difference in accuracy of the single redshift inference was negligible. As a result, we do not transform redshifts throughout this chapter.

A plot of this prior distribution, fitted to the simulated LSST-like training data described in section 7.3 and plotted using `corner.py` (Foreman-Mackey, 2016), is shown in Figure 7.2. The ability to plot this distribution is an advantage to this GMM method. As described above, machine learning methods can act as black boxes, where

what has been learned is a complicated function approximator that can be difficult to interpret. In contrast, the central object being learned here is the joint flux-redshift distribution, a meaningful statistical object that can be plotted, sampled from and manipulated mathematically.

7.1.2 Utilising blended training data

The derivations detailed in sections 7.2.1 and 7.2.2 are presented for a scalar redshift z. However, it should be noted that these single-constituent results also hold for a vector $\boldsymbol{z}$. As a result, this method can be generalised so that the model is fitted to blended training data, i.e., a vector of blended fluxes with the associated vector of redshifts for each constituent.

Utilising blended training data would allow the method to infer both redshifts and the number of constituents accurately in cases where the blended constituents were systematically different from non-blended constituents. The cost of this, however, is an increase in the required size of the training set. Machine learning-based methods require a training set that is representative of the test set in order to be accurate. A blended training set would therefore have to contain sufficient examples of all possible pairs of constituents, rather than the constituents alone as required for the results in sections 7.2.3 and 7.2.4.

7.1.3 Cross-validating the number of mixture components

The procedure described in section 7.1.1 will fit the weights, means and covariances of the GMM for a fixed number of components. However, it is difficult *a priori* to choose this number; including more components within the mixture allows it more flexibility, but too many will cause the model to overfit. Given enough mixture components, the variance of each component will approach zero, with each being responsible for only a single sample. While this will significantly increase the likelihood of the training set, it will also cause the model to generalise extremely poorly.

Overfitting is a general concern when fitting machine learning models. As a result, various techniques for preventing overfitting have been suggested. These include restricting the dimensionality of the parameter space as we do here by fixing the number of components, disfavouring overfitted parameters through regularisation (e.g., Hoerl and Kennard, 1970) or Bayesian priors (e.g., MacKay, 1992), and stopping training before overfitting occurs (e.g., Prechelt, 1998).

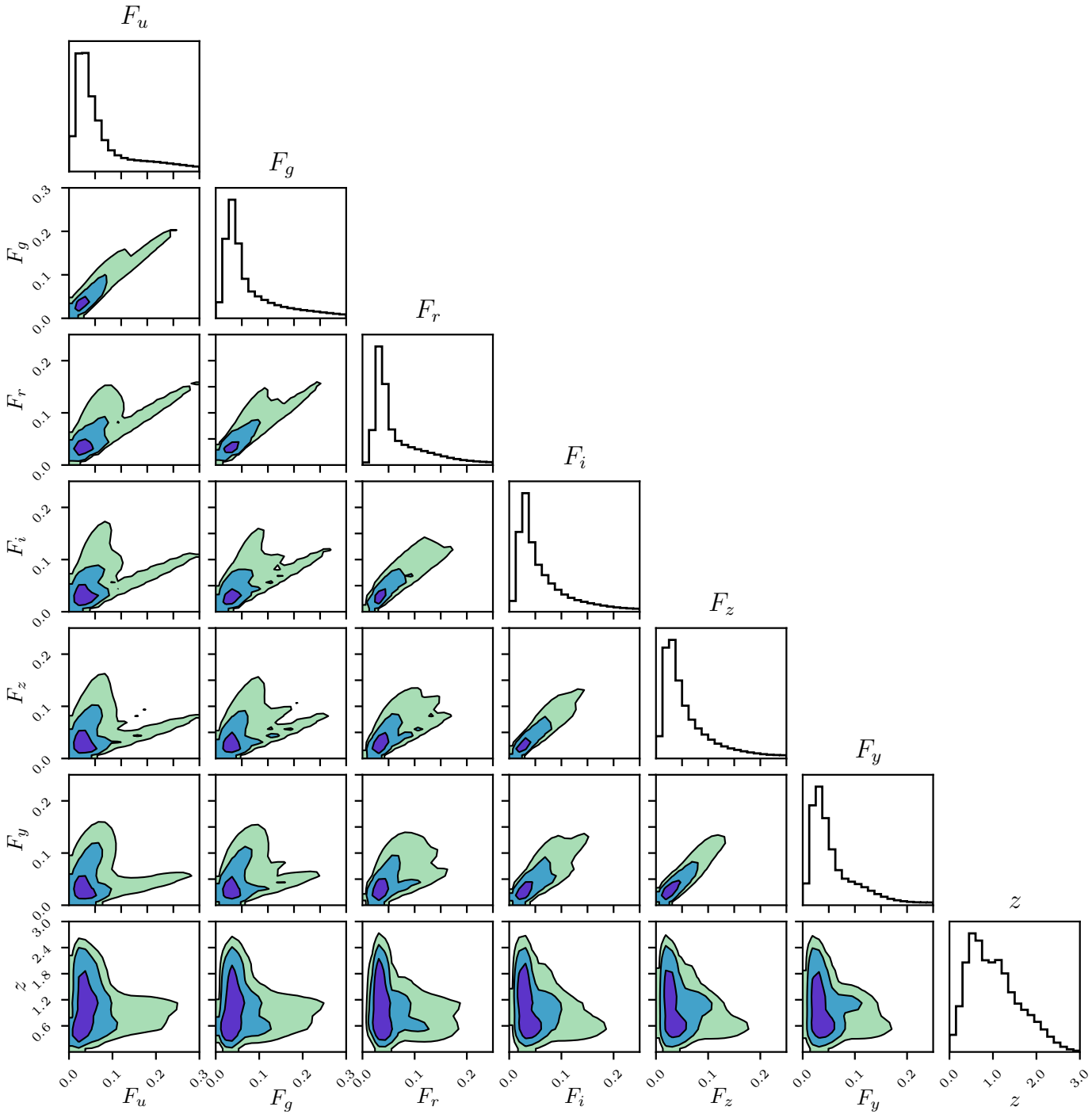

Figure 7.2: Corner plot of an example flux-redshift distribution fitted by our model. This density shown here is visualised using 10^6 samples drawn from a model that was fitted to the LSST-like simulations presented in section 7.3.

The ability for a machine learning method to generalise and whether it has been overfitted can be tested by using a a validation set, an additional set of data where the input and output are known but is not used during the training. By measuring the difference between the prediction and the known ground truth, the model can be evaluated.

It is useful to point out that a corollary to the notion of overfitting is that the fitting procedure need not converge to a global maximum, as that set of parameters will overfit the data. Instead, local maxima can be nearly as accurate on the test set, while generalising much better (Choromanska et al., 2014). Therefore, it is reasonable to use parameters corresponding to local maxima that are found to perform well during validation. This can avoid expending significant optimisation effort attempting to fit the global maximum.

To choose the number of components, we use k-fold cross validation, a method that repeatedly splits the data into training and validation sets. The training set is first split into k subsets. The model is then trained on $k-1$ subsets of this data, assuming a fixed number of mixture components M. The remaining subset is then used for validation. By evaluating the model using the fluxes of this subset, the redshift predictions can be compared to the known truth and scored based on their accuracy. This training and validation is repeated k times for each number of components considered so that each subset is used for evaluation once. The average score can then be used to evaluate each number of components.

To evaluate the accuracy of the redshift predictions, we use the RMS scatter. Given a predicted redshift $z_{\mathrm{p},g}$ and a spectroscopic redshift $\hat{z}_{\mathrm{s},g}$ for galaxy g, the normalised error is defined as

$$\tilde{\delta}_g = \frac{\hat{z}_{\mathrm{s},g} - z_{\mathrm{p},g}}{1 + \hat{z}_{\mathrm{s},g}} \,. \tag{7.9}$$

After calculating this error for n_{g} galaxies, the RMS scatter for the sample is then given by

$$\sigma_{\mathrm{RMS}} = \sqrt{\frac{1}{n_{\mathrm{g}}} \sum_g \tilde{\delta}_g^2} \,. \tag{7.10}$$

This metric is evaluated using k-fold validation for each number of mixture components M being considered. We then choose M to be the number of components that minimises the RMS scatter averaged over each of the k folds.

7.1.4 Sampling from Gaussian mixture models

One of the significant advantages of using GMMs is that they can be efficiently sampled from without using methods such as MCMC. Since they are simply linear combinations of component distributions, a simple sampling scheme is to randomly select one of the components with a probability given by the weights, and then to draw a sample from the respective multivariate Gaussian.

This sampling scheme allows GMMs to be sampled efficiently and without rejection. However, the addition of the boundary prior described in section 7.2.1 means that samples with negative fluxes and redshifts are rejected during inference. Nevertheless, the efficiency of this sampling scheme means that this does not pose a problem, since many samples can still be drawn from the relevant posterior with little computational effort.

7.1.5 Compressed storage of PDFs

As described above, it is important that the results of photometric redshifts are represented as a PDF. However, given the large sample sizes of future galaxy surveys like LSST, storing these PDFs can present a problem. While a point estimate of the redshift and an associated error can be stored simply as two real numbers, a PDF will generally require many more. A naive representation of this distribution is a histogram where the redshift bins are fixed for all sources. While this is simple, it is not space efficient.

This problem was first investigated by Carrasco Kind and Brunner (2014), who proposed a sparse basis representation using Gaussian and Voigt distributions. Using this method, the PDF can be stored in a single signed integer per basis function, with $\mathcal{O}(10)$ basis functions required to accurately reconstruct the original PDFs. Malz et al. (2018) test PDF compression methods by measuring the Kullback-Leibler divergence between the original and compressed PDFs. They suggest storing the redshifts corresponding to equally-spaced quantiles as an alternative to histograms.

The posteriors presented here are GMMs, potentially multiplied by an additional physical constraint. This representation permits a simple compression technique of discarding low-weight components. By construction, the number of components in the mixture describing the prior is the same as the mixture describing the redshift posterior. However, the latter is generally significantly more compact, describing the density over the parameter space for a single source only, rather than the entire population. It is

therefore reasonable to expect that this posterior distribution could be represented by fewer components than the prior.

If additional computation can be afforded for a further reduction in storage space, mixture components can also be merged into a smaller number of approximating components. This procedure is known as mixture reduction (see, e.g., West, 1993; Williams and Maybeck, 2006; Runnalls, 2007; Schieferdecker and Huber, 2009).

7.2 Deriving posteriors and evidences

7.2.1 Single-constituent posterior

We now derive the posterior distribution assuming that the source consists of a single, unblended constituent galaxy. The redshift under this model can then be inferred by sampling from this posterior, as described in section 7.1.4. We start by marginalising over the true, latent flux vector $\boldsymbol{F}$, giving

$$P(z \mid \hat{\boldsymbol{F}}) = \int P(z, \boldsymbol{F} \mid \hat{\boldsymbol{F}}) \, \mathrm{d}\boldsymbol{F} \,. \tag{7.11}$$

Applying Bayes rule, this becomes

$$P(z \mid \hat{\boldsymbol{F}}) \propto \int P(\hat{\boldsymbol{F}} \mid \boldsymbol{F}) \; P(z, \boldsymbol{F}) \, \mathrm{d}\boldsymbol{F} \,, \tag{7.12}$$

where the unnecessary redshift conditioning has been dropped from the likelihood. We assume the likelihood to be a multivariate Gaussian centred on the observed fluxes, i.e.,

$$P(\hat{\boldsymbol{F}} \mid \boldsymbol{F}) = \mathcal{N}(\boldsymbol{F} \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}^{\hat{F}}) \,, \tag{7.13}$$

where $\underline{\boldsymbol{\Sigma}}^{\hat{F}}$ is the covariance matrix of the observation. Galaxy surveys typically assume the errors on observed fluxes in each band to be independent, i.e., given as a flux and an error. In this case, the covariance matrix would simply be diagonal. No assumption is made about this covariance throughout however, allowing fluxes to be correlated in general.

The prior in equation 7.12 is given by the GMM described above. This prior is the only term involving the redshift; it fully represents the relation between flux and redshift learned from the training set.

Inserting both the prior and the likelihood into equation 7.12, the posterior becomes

$$P(z \mid \hat{\boldsymbol{F}}) \propto \sum_k w^k \int \mathcal{N}(\boldsymbol{F} \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}^{\hat{F}})\, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k)\, \mathrm{d}\boldsymbol{F}\,. \tag{7.14}$$

This posterior now contains the product of two Gaussian PDFs, albeit with different dimensionalities. We proceed by combining these two densities into a single multivariate Gaussian. This is analogous to the derivation of Bovy et al. (2012). However, as described above, we do not make use of the convolution property of multivariate Gaussians, instead forming the product explicitly. To do this, we write our posterior in terms of a parameter vector $\boldsymbol{\theta}$ partitioned into redshift and fluxes, i.e.,

$$\boldsymbol{\theta} = \begin{pmatrix} z \\ \\ \boldsymbol{F} \end{pmatrix}\,. \tag{7.15}$$

Throughout, we label the redshift and flux blocks of parameters partitioned in the same way with z and f respectively.

The likelihood involves only the flux partition of the parameter vector. However, our prior has support over both redshift and flux, i.e., all of $\boldsymbol{\theta}$. The component parameters are thus partitioned in the same way so that the mean and covariance are given by

$$\boldsymbol{\mu}^k = \begin{pmatrix} \mu_{\mathrm{z}}^k \\ \\ \boldsymbol{\mu}_{\mathrm{f}}^k \end{pmatrix} \tag{7.16}$$

and

$$\underline{\boldsymbol{\Sigma}}^k = \begin{pmatrix} \Sigma_{\mathrm{zz}}^k & \boldsymbol{\Sigma}_{\mathrm{zf}}^k \\ \\ \boldsymbol{\Sigma}_{\mathrm{fz}}^k & \underline{\boldsymbol{\Sigma}}_{\mathrm{ff}}^k \end{pmatrix} \tag{7.17}$$

respectively. The product of these two densities is most easily written in terms of the natural parametrisation[1] of the multivariate Gaussian. This has a density given by

$$\tilde{\mathcal{N}}(\boldsymbol{x}|\boldsymbol{\eta}, \underline{\boldsymbol{\Lambda}}) = \exp\left[\alpha + \boldsymbol{\eta}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\underline{\boldsymbol{\Lambda}}\boldsymbol{x}\right]\,, \tag{7.18}$$

where we have added a tilde to notate the alternative parametrisation. The normalisation factor is given by

$$\alpha = -\frac{1}{2}\left[d\log(2\pi) - \log|\underline{\boldsymbol{\Lambda}}| + \boldsymbol{\eta}^T\underline{\boldsymbol{\Lambda}}^{-1}\boldsymbol{\eta}\right]\,, \tag{7.19}$$

[1]This is also referred to as the canonical or information parametrisation.

and the covariance matrix and mean vector are replaced with the natural parameters $\underline{\boldsymbol{\Lambda}} \equiv \underline{\boldsymbol{\Sigma}}^{-1}$ and $\boldsymbol{\eta} \equiv \underline{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\mu}$. respectively. The inverse covariance matrix $\underline{\boldsymbol{\Lambda}}$ is known as the precision matrix. The product of the two densities in equation 7.14 can then be combined into a single multivariate Gaussian written in this natural parametrisation, given by

$$\mathcal{N}(\boldsymbol{F} \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}^{\hat{F}})\, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) = c_1^k \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}})\,, \tag{7.20}$$

where the new parameters are

$$\underline{\boldsymbol{\Lambda}}^{k\hat{F}} = (\underline{\boldsymbol{\Sigma}}^k)^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & (\underline{\boldsymbol{\Sigma}}^{\hat{F}})^{-1} \end{pmatrix} \tag{7.21}$$

and

$$\boldsymbol{\eta}^{k\hat{F}} = (\underline{\boldsymbol{\Sigma}}^k)^{-1}\boldsymbol{\mu}^k + \begin{pmatrix} 0 \\ (\underline{\boldsymbol{\Sigma}}^{\hat{F}})^{-1}\hat{\boldsymbol{F}} \end{pmatrix}. \tag{7.22}$$

Conveniently, the constant of proportionality c can also be written in terms of a multivariate Gaussian in standard parametrisation. This is given by

$$c_1^k = \mathcal{N}(\boldsymbol{\mu}_{\mathrm{f}}^k \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}_{\mathrm{ff}}^k + \underline{\boldsymbol{\Sigma}}^{\hat{F}})\,. \tag{7.23}$$

These results are close to a standard property (e.g., Petersen and Pedersen, 2014) where the product of two multivariate Gaussian densities is also a multivariate Gaussian. However, the differing dimensionalities of the two densities in equation 7.20 slightly alter the expressions for the new parameters.

Inserting these results into equation 7.14 and moving constant terms outside of the integral, the expression for the posterior becomes

$$P(z \mid \hat{\boldsymbol{F}}) \propto \sum_k w^k c_1^k \int \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}})\, \mathrm{d}\boldsymbol{F}\,. \tag{7.24}$$

In principle, this integral can be done analytically by moving back to standard parametrisation, i.e., $\underline{\boldsymbol{\Sigma}}^{k\hat{F}} = \left(\underline{\boldsymbol{\Lambda}}^{k\hat{F}}\right)^{-1}$ and $\boldsymbol{\mu}^{k\hat{F}} = \underline{\boldsymbol{\Sigma}}^{k\hat{F}}\boldsymbol{\eta}^{k\hat{F}}$. The marginalisation can then be done by dropping the corresponding elements from the mean vector and covariance matrix, giving

$$P(z \mid \hat{\boldsymbol{F}}) \propto \sum_k w^k c_1^k \mathcal{N}(z \mid \boldsymbol{\mu}_{\mathrm{z}}^{k\hat{F}}, \underline{\boldsymbol{\Sigma}}_{\mathrm{zz}}^{k\hat{F}})\,. \tag{7.25}$$

Note that this is simply a one-dimensional Gaussian mixture model with a new set of weights given by $w^{k\hat{F}} \equiv w^k c_1^k$.

An important caveat to this result, however, is that the limits of integration are assumed to be $(-\infty, \infty)$; that is, non-physical negative fluxes contribute to the integral. This is the same assumption as used in the derivation in Bovy et al. (2012) using the convolution property of multivariate Gaussians. For this non-blended photo-z, this assumption is sound since the latent fluxes are strongly constrained by the likelihood, meaning that negative fluxes will be strongly down-weighted. However, this will not be the case for the blended photo-z derived in section 7.2.3 where only the sum of two latent flux vectors is observed.

An alternative approach is to add the boundary prior $\psi(z, \boldsymbol{F})$ as described in section 7.1.1. This has two effects. Firstly, the prior with this addition must be explicitly normalised, a necessary condition for the model selection. The normalisation factor is given by an integral over the unnormalised prior, i.e.,

$$\mathcal{A}_1 = \iint \psi(z, \boldsymbol{F}) \sum_k w^k \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}^k, \underline{\boldsymbol{\Sigma}}^k) \, \mathrm{d}z \, \mathrm{d}\boldsymbol{F} \,. \tag{7.26}$$

This integral can be efficiently estimated using Monte Carlo integration. First, a set of redshifts and fluxes $\{z, \boldsymbol{F}\}$ is sampled from the mixture, as described in section 7.1.4. Since the prior without the boundary prior is normalised to unity as in equation 7.2, this integral is then equal to fraction of these samples obeying the boundary prior, i.e., where $\psi(z, \boldsymbol{F}) = 1$.

The second effect of adding the boundary prior is that marginalising over fluxes is no longer analytic. Inserting the boundary prior and the corresponding prior normalisation $\mathcal{A}_1$, the posterior we want to sample from is given by

$$P(z \mid \hat{\boldsymbol{F}}) \propto \mathcal{A}_1 \sum_k w^k c_1^k \int \psi(z, \boldsymbol{F}) \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}}) \, \mathrm{d}\boldsymbol{F} \,. \tag{7.27}$$

However, the boundary prior makes this integral non-analytic and the resulting posterior is not a standard GMM, meaning that it cannot be sampled as described in section 7.1.4. Instead, we sample from the density given by

$$P(z, \boldsymbol{F} \mid \hat{\boldsymbol{F}}) \propto \mathcal{A}_1 \sum_k w^k c_1^k \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}}) \,. \tag{7.28}$$

This is the desired posterior from equation 7.27 without the marginalisation over fluxes and where we have neglected the boundary prior term. This can then be corrected for by rejecting any sample that contains negative fluxes or redshift, leaving only the samples

that obey the boundary prior. The marginalisation can then be done trivially by discarding the fluxes and considering only the redshift part of the remaining samples. Since equation 7.28 is simply a new Gaussian mixture model as before, sampling from this distribution is extremely computationally efficient, as detailed in section 7.1.4. As described above, the inclusion of the boundary prior is most important for the blended photo-z, though we include it here for completeness and consistency with the blended case later.

7.2.2 Single-constituent evidence

One of the more computationally demanding aspects of the method of chapter 6 is the use of nested sampling in order to calculate the evidence. A significant advantage of the GMM method presented here is that this expensive integral can be evaluated much more quickly, an important feature for applying the method to future surveys.

The single-constituent evidence $\mathcal{E}^1$ is defined to be the integral of the unnormalised posterior over the full parameter space, i.e.,

$$\mathcal{E}^1 = \int\int P(\hat{\boldsymbol{F}} \mid \boldsymbol{F})\; P(z, \boldsymbol{F})\; \mathrm{d}\boldsymbol{F}\; \mathrm{d}z\,. \tag{7.29}$$

As described above, by ignoring the boundary prior, the integral over fluxes can be performed analytically to give a new Gaussian mixture model. Inserting this result into the evidence integral, equation 7.29 becomes

$$\mathcal{E}^1 = \sum_k w^k\; c_1^k \int \mathcal{N}(z \mid \boldsymbol{\mu}_{\mathrm{z}}^{k\hat{F}}, \underline{\boldsymbol{\Sigma}}_{\mathrm{zz}}^{k\hat{F}})\; \mathrm{d}z\,. \tag{7.30}$$

Since the multivariate Gaussian density of each component is normalised to unity, the evidence is then given simply by the sum over the new mixture weights, i.e.,

$$\mathcal{E}^1 = \sum_k w^k c_1^k \equiv \sum_k w^{k\hat{F}}\,. \tag{7.31}$$

In this case, the evidence is analytic and therefore easy to compute. However, as above, computing these integrals analytically implicitly involves contributions from non-physical negative fluxes and redshifts.

To combat this, we can numerically integrate the non-marginalised posterior of fluxes and redshifts including the boundary prior introduced in section 7.1.1 and the

accompanying normalisation from equation 7.26, i.e.,

$$\mathcal{E}^1 = \iint \mathcal{A}_1 \sum_k w^k c_1^k \psi(z, \boldsymbol{F}) \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}}) \, \mathrm{d}\boldsymbol{F} \, \mathrm{d}z \,. \tag{7.32}$$

This integral can be evaluated numerically by using fluxes and redshifts sampled from the non-marginalised posterior with the boundary prior removed, given in equation 7.28. This is another Gaussian mixture model, and thus these samples are computationally efficient to draw, as described in section 7.1.4. In addition, the posterior samples drawn for inference are also sampled from equation 7.28 and so can be reused here, saving computation.

Given a set of samples $\{z, \boldsymbol{F}\}$ from equation 7.28, only a fraction $\mathcal{F}_1$ of these will contain no negative fluxes. Unlike equation 7.26, however, this density is not normalised to unity, but rather

$$\begin{aligned} \mathcal{V}_1 &\equiv \mathcal{A}_1 \sum_k w^k c_1^k \iint \tilde{\mathcal{N}}(z, \boldsymbol{F} \mid \boldsymbol{\eta}^{k\hat{F}}, \underline{\boldsymbol{\Lambda}}^{k\hat{F}}) \, \mathrm{d}z \, \mathrm{d}\boldsymbol{F} \\ &= \mathcal{A}_1 \sum_k w^k c_1^k \,. \end{aligned} \tag{7.33}$$

By using this to compute a Monte Carlo estimate of the integral, the evidence can therefore be estimated to be

$$\mathcal{E}^1 \approx \mathcal{V}_1 \mathcal{F}_1 = \mathcal{A}_1 \mathcal{F}_1 \sum_k w^k c_1^k \equiv \mathcal{A}_1 \mathcal{F}_1 \sum_k w^{k\hat{F}} \,. \tag{7.34}$$

7.2.3 Two-constituent posterior

We now extend the inference method to the case of a blended source consisting of two constituent galaxies by deriving the two-constituent posterior. Here, the parameters we wish to infer are the redshifts of each constituent $\{z\} = \{z_1, z_2\}$, given the data vector of observed fluxes $\hat{\boldsymbol{F}}$.

As before, we start by marginalising over the latent flux vectors. As this is the two-constituent posterior, there are now two flux vectors to marginalise over, $\{\boldsymbol{F}\} = \{\boldsymbol{F}_1, \boldsymbol{F}_2\}$, one for each galaxy. The posterior is therefore given by

$$P(\{z\} \mid \hat{\boldsymbol{F}}) = \int P(\{z\}, \{\boldsymbol{F}\} \mid \hat{\boldsymbol{F}}) \, \mathrm{d}\{\boldsymbol{F}\} \,. \tag{7.35}$$

Applying Bayes rule, this becomes

$$P(\{z\} \mid \hat{\boldsymbol{F}}) \propto \int P(\hat{\boldsymbol{F}} \mid \{\boldsymbol{F}\})\, P(\{z\}, \{\boldsymbol{F}\})\, \mathrm{d}\{\boldsymbol{F}\}\,, \tag{7.36}$$

where $P(\{z\}, \{\boldsymbol{F}\})$ is the joint prior over flux and redshift for both constituents. This prior can be factorised to be written in terms of the individual constituent priors $P(z, \boldsymbol{F})$, allowing the GMM to be inserted. However, as described in section 6.1, the parameters of each constituent are correlated. Thus, the joint prior can be written as

$$P(\{z\}, \{\boldsymbol{F}\}) \propto P(z_1, \boldsymbol{F}_1) P(z_2, \boldsymbol{F}_2) M(z_1, z_2)\,, \tag{7.37}$$

where the blending-related correlations have been factored into a single term

$$M(z_1, z_2) = \pi(z_1, z_2)\, [1 + \xi(z_1, z_2)]\,. \tag{7.38}$$

Here, $\xi(z_1, z_2)$ is the two-point galaxy correlation function, evaluated at the line-of-sight comoving distance between z_2 and z_1. This correlation function is commonly modelled as a power law (e.g., Peebles, 2001). However, we make no assumption of its form throughout this derivation, requiring only that it can be evaluated given a pair of redshifts. This correlation function was found to have little effect in chapter 6 so the results throughout assume $\xi(z_1, z_2) = 0$. Nevertheless, we include it in the derivations here for completeness. The term $\pi(z_1, z_2)$ represents the sorting condition, given by

$$\pi(z_1, z_2) = \begin{cases} 1 & \text{for } z_1 \leq z_2 \\ 0 & \text{otherwise.} \end{cases} \tag{7.39}$$

The need for these terms is discussed in detail in chapter 6.

Any selection effects on the training set are already captured in the prior through the training step. This assumes that the training set is sufficiently representative of the test set, though we note that this caveat applies to machine learning methods in general. The selection effect term of chapter 6 simply acts to disfavour inferring fluxes such that the total flux is near the survey limit, as they are *a priori* less likely to have been selected. Since the total flux is well constrained by observations, this term has little effect on parameter inferences. Instead, its use is motivated by making the magnitude prior proper. This is necessary for evaluating the marginal likelihood for model comparison. However, our GMM prior is proper by construction. As a result, we do not include the selection effect term here.

As in section 7.2.1, the model selection requires the joint prior to be normalised.

We do this by integrating the prior using Monte Carlo integration. To be able to draw samples from the prior efficiently, we insert the definitions of each term and combine into another Gaussian mixture that can be sampled as described in section 7.1.4. We also include the boundary prior described in section 7.1.1 in each constituent prior to prevent contributions to the density from non-physical negative fluxes and redshifts.

Inserting the GMM, correlation and boundary prior terms into equation 7.37, the joint prior becomes

$$\begin{aligned} P(\{z\},\{\hat{\boldsymbol{F}}\}) \propto M(z_1,z_2)\psi(z_1,\boldsymbol{F}_1)\psi(z_2,\boldsymbol{F}_2)\sum_k\sum_j w^k w^j \times \\ \mathcal{N}(z_1,\boldsymbol{F}_1 \mid \boldsymbol{\mu}^k,\underline{\boldsymbol{\Sigma}}^k)\mathcal{N}(z_2,\boldsymbol{F}_2 \mid \boldsymbol{\mu}^j,\underline{\boldsymbol{\Sigma}}^j)\,. \end{aligned} \tag{7.40}$$

We now follow an analogous method to that of section 7.2.1 by combining the two multivariate Gaussians into a single density. We start by defining a partitioned parameter vector that each density can be written in terms of. This is given by

$$\boldsymbol{\psi} = \begin{pmatrix} z_1 \\ \boldsymbol{F}_1 \\ z_2 \\ \boldsymbol{F}_2 \end{pmatrix}. \tag{7.41}$$

The product of the densities in equation 7.40 can then be written as a single Gaussian density in terms of this parameter vector

$$\mathcal{N}(z_1,\boldsymbol{F}_1 \mid \boldsymbol{\mu}^k,\underline{\boldsymbol{\Sigma}}^k)\,\mathcal{N}(z_2,\boldsymbol{F}_2 \mid \boldsymbol{\mu}^j,\underline{\boldsymbol{\Sigma}}^j) \;= \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj},\underline{\boldsymbol{\Sigma}}^{kj})\,, \tag{7.42}$$

where the new mean vector is given by

$$\boldsymbol{\mu}^{kj} = \begin{pmatrix} \boldsymbol{\mu}^k \\ \boldsymbol{\mu}^j \end{pmatrix} = \begin{pmatrix} \mu_{\mathrm{z}}^k \\ \boldsymbol{\mu}_{\mathrm{f}}^k \\ \mu_{\mathrm{z}}^j \\ \boldsymbol{\mu}_{\mathrm{f}}^j \end{pmatrix} \tag{7.43}$$

and the covariance matrix

$$\underline{\boldsymbol{\Sigma}}^{kj} = \begin{pmatrix} \underline{\boldsymbol{\Sigma}}^{k} & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\boldsymbol{\Sigma}}^{j} \end{pmatrix} = \begin{pmatrix} \Sigma^{k}_{\mathrm{zz}} & \boldsymbol{\Sigma}^{k}_{\mathrm{zf}} & 0 & \mathbf{0} \\ \boldsymbol{\Sigma}^{k}_{\mathrm{fz}} & \underline{\boldsymbol{\Sigma}}^{k}_{\mathrm{ff}} & \mathbf{0} & \underline{\mathbf{0}} \\ 0 & \mathbf{0} & \Sigma^{j}_{\mathrm{zz}} & \boldsymbol{\Sigma}^{j}_{\mathrm{zf}} \\ \mathbf{0} & \underline{\mathbf{0}} & \boldsymbol{\Sigma}^{j}_{\mathrm{fz}} & \underline{\boldsymbol{\Sigma}}^{j}_{\mathrm{ff}} \end{pmatrix} . \tag{7.44}$$

This combination is trivial since we assume that all correlations between the two constituents have already been factored out into $M(\{z\}, \{\boldsymbol{F}\})$. As a result, the two constituent priors are independent and can be combined with the block diagonal covariance matrix defined in equation 7.44. The joint prior thus becomes

$$\begin{aligned} P(\{z\}, \{\hat{\boldsymbol{F}}\}) \propto M(z_1, z_2)\psi(z_1, \boldsymbol{F}_1)\psi(z_2, \boldsymbol{F}_2)\times \\ \sum_k \sum_j w^k w^j \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj}, \underline{\boldsymbol{\Sigma}}^{kj}) \,, \end{aligned} \tag{7.45}$$

i.e., a GMM multiplied by several additional terms. The normalisation of this prior is then given by the integral

$$\begin{aligned} \mathcal{A}_2 = \iiiint M(z_1, z_2)\psi(z_1, \boldsymbol{F}_1)\psi(z_2, \boldsymbol{F}_2)\times \\ \sum_k \sum_j w^k w^j \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj}, \underline{\boldsymbol{\Sigma}}^{kj}) \, \mathrm{d}z_1 \, \mathrm{d}z_2 \, \mathrm{d}\hat{\boldsymbol{F}}_\alpha \, \mathrm{d}\hat{\boldsymbol{F}}_\beta \,. \end{aligned} \tag{7.46}$$

Analogously to equation 7.26, this can be evaluated using samples drawn from the Gaussian mixture, i.e.,

$$\{z_1, z_2, \boldsymbol{F}_1, \boldsymbol{F}_2\} \sim G(\boldsymbol{\psi}) = \sum_k \sum_j w^k w^j \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj}, \underline{\boldsymbol{\Sigma}}^{kj}) \,. \tag{7.47}$$

Given $n_{\mathcal{A}}$ of these samples $\{z_1^i, z_2^i, \boldsymbol{F}_1^i, \boldsymbol{F}_2^i \mid i = 1 \ldots n_{\mathcal{A}}\}$, we can compute a Monte Carlo integration of $\mathcal{A}_2$ through importance sampling. Since $G(\boldsymbol{\psi})$ is normalised to unity, this integral is given by

$$\mathcal{A}_2 = \sum_i \frac{[1 + \xi(z_1^i, z_2^i)] \, \pi(z_1^i, z_2^i)\psi(z_1^i, \boldsymbol{F}_1^i)\psi(z_2^i, \boldsymbol{F}_2^i)}{n_{\mathcal{A}}} \,. \tag{7.48}$$

If the correlation function and sorting condition were ignored, this would simply be equal to the fraction of samples that obey the boundary prior, as in the definition of

$\mathcal{A}_1$. Thus, the joint prior is given by

$$P(\{z\},\{\hat{\boldsymbol{F}}\}) = \mathcal{A}_2 M(z_1,z_2)\psi(z_1,\boldsymbol{F}_1)\psi(z_2,\boldsymbol{F}_2)\times \sum_k \sum_j w^k w^j \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj}, \underline{\boldsymbol{\Sigma}}^{kj})\,. \tag{7.49}$$

This joint prior can then be inserted into equation 7.36 alongside the definition of the likelihood to develop the posterior. As before, we assume that the likelihood is a multivariate Gaussian centred on the observed fluxes, though we now model the flux as the sum of the constituent fluxes, i.e.,

$$P(\hat{\boldsymbol{F}} \mid \boldsymbol{F}) = \mathcal{N}(\boldsymbol{F}_1 + \boldsymbol{F}_2 \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}^{\hat{F}})\,. \tag{7.50}$$

Inserting this likelihood and the joint prior into equation 7.36, the posterior becomes

$$\begin{aligned} P(\{z\} \mid \hat{\boldsymbol{F}}) \propto & \mathcal{A}_2 \iint M(z_1,z_2)\psi(z_1,\boldsymbol{F}_1)\psi(z_2,\boldsymbol{F}_2)\times \\ & \sum_k \sum_j w^k w^j \mathcal{N}(\boldsymbol{F}_1 + \boldsymbol{F}_2 \mid \hat{\boldsymbol{F}}, \underline{\boldsymbol{\Sigma}}^{\hat{F}})\times \\ & \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\mu}^{kj}, \underline{\boldsymbol{\Sigma}}^{kj})\, \mathrm{d}\boldsymbol{F}_1\, \mathrm{d}\boldsymbol{F}_2\,. \end{aligned} \tag{7.51}$$

To combine the prior term with the likelihood, we rewrite it in terms of natural parameters partitioned in the same way as equation 7.41. These new parameters are given by

$$\boldsymbol{\eta}^{kj} = \begin{pmatrix} \boldsymbol{\eta}^k \\ \boldsymbol{\eta}^j \end{pmatrix} = \begin{pmatrix} \eta_{\mathrm{z}}^k \\ \boldsymbol{\eta}_{\mathrm{f}}^k \\ \eta_{\mathrm{z}}^j \\ \boldsymbol{\eta}_{\mathrm{f}}^j \end{pmatrix} \tag{7.52}$$

and

$$\underline{\boldsymbol{\Lambda}}^{kj} = \begin{pmatrix} \underline{\boldsymbol{\Lambda}}^k & \underline{\mathbf{0}} \\ \underline{\mathbf{0}} & \underline{\boldsymbol{\Lambda}}^j \end{pmatrix} = \begin{pmatrix} \Lambda_{\mathrm{zz}}^k & \boldsymbol{\Lambda}_{\mathrm{zf}}^k & 0 & \mathbf{0} \\ \boldsymbol{\Lambda}_{\mathrm{fz}}^k & \underline{\boldsymbol{\Lambda}}_{\mathrm{ff}}^k & \mathbf{0} & \underline{\mathbf{0}} \\ 0 & \mathbf{0} & \Lambda_{\mathrm{zz}}^j & \boldsymbol{\Lambda}_{\mathrm{zf}}^j \\ \mathbf{0} & \underline{\mathbf{0}} & \boldsymbol{\Lambda}_{\mathrm{fz}}^j & \underline{\boldsymbol{\Lambda}}_{\mathrm{ff}}^j \end{pmatrix}. \tag{7.53}$$

By also rewriting the likelihood in terms of the natural parameters $\underline{\boldsymbol{\Lambda}}^{\hat{F}} \equiv \left(\underline{\boldsymbol{\Sigma}}^{\hat{F}}\right)^{-1}$ and

$\boldsymbol{\eta}^{\hat{F}} \equiv \underline{\boldsymbol{\Lambda}}^{\hat{F}} \hat{\boldsymbol{F}}$, the posterior becomes

$$\begin{aligned} P(\{z\} \mid \hat{\boldsymbol{F}}) \propto \iint & \mathcal{A}_2 M(z_1, z_2) \psi(z_1, \boldsymbol{F}_1) \psi(z_2, \boldsymbol{F}_2) \times \\ & \sum_k \sum_j w^k w^j \tilde{\mathcal{N}}(\boldsymbol{F}_1 + \boldsymbol{F}_2 \mid \boldsymbol{\eta}^{\hat{F}}, \underline{\boldsymbol{\Lambda}}^{\hat{F}}) \times \\ & \tilde{\mathcal{N}}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj}, \underline{\boldsymbol{\Lambda}}^{kj}) \, \mathrm{d}\boldsymbol{F}_1 \, \mathrm{d}\boldsymbol{F}_2 \, . \end{aligned} \tag{7.54}$$

The two remaining densities can now be combined into a single term given by

$$\tilde{\mathcal{N}}(\boldsymbol{F}_1 + \boldsymbol{F}_2 \mid \boldsymbol{\eta}^{\hat{F}}, \underline{\boldsymbol{\Lambda}}^{\hat{F}}) \, \tilde{\mathcal{N}}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj}, \underline{\boldsymbol{\Lambda}}^{kj}) \propto \tilde{\mathcal{N}}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj\hat{F}}, \underline{\boldsymbol{\Lambda}}^{kj\hat{F}}) \, , \tag{7.55}$$

where the combined parameters are given by

$$\boldsymbol{\eta}^{kj\hat{F}} = \begin{pmatrix} \eta_{\mathrm{z}}^k \\ \boldsymbol{\eta}_{\mathrm{f}}^k + \boldsymbol{\eta}^{\hat{F}} \\ \eta_{\mathrm{z}}^j \\ \boldsymbol{\eta}_{\mathrm{f}}^j + \boldsymbol{\eta}^{\hat{F}} \end{pmatrix} \tag{7.56}$$

and

$$\underline{\boldsymbol{\Lambda}}^{kj\hat{F}} = \begin{pmatrix} \Lambda_{\mathrm{zz}}^k & \boldsymbol{\Lambda}_{\mathrm{zf}}^k & 0 & \mathbf{0} \\ \boldsymbol{\Lambda}_{\mathrm{fz}}^k & \underline{\boldsymbol{\Lambda}}_{\mathrm{ff}}^k + \underline{\boldsymbol{\Lambda}}^{\hat{F}} & \mathbf{0} & \underline{\boldsymbol{\Lambda}}^{\hat{F}} \\ 0 & \mathbf{0} & \Lambda_{\mathrm{zz}}^j & \boldsymbol{\Lambda}_{\mathrm{zf}}^j \\ \mathbf{0} & \underline{\boldsymbol{\Lambda}}^{\hat{F}} & \boldsymbol{\Lambda}_{\mathrm{fz}}^j & \underline{\boldsymbol{\Lambda}}_{\mathrm{ff}}^j + \underline{\boldsymbol{\Lambda}}^{\hat{F}} \end{pmatrix} . \tag{7.57}$$

As before, the constant of proportionality c_2^{kj} in equation 7.55 can also be written in terms of another multivariate Gaussian density

$$c_2^{kj} = \mathcal{N}\left(\mu_{\mathrm{f}}^k + \mu_{\mathrm{f}}^j \mid \hat{\boldsymbol{F}}, \left[\Sigma^{\hat{F}} + \Sigma_{\mathrm{ff}}^k + \Sigma_{\mathrm{ff}}^j\right]\right) . \tag{7.58}$$

The posterior is thus given by

$$\begin{aligned} P(\{z\} \mid \hat{\boldsymbol{F}}) \propto \iint & \mathcal{A}_2 M(z_1, z_2) \psi(z_1, \boldsymbol{F}_1) \psi(z_2, \boldsymbol{F}_2) \times \\ & \sum_k \sum_j w^k w^j c_2^{kj} \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj\hat{F}}, \underline{\boldsymbol{\Lambda}}^{kj\hat{F}}) \, \mathrm{d}\boldsymbol{F}_1 \, \mathrm{d}\boldsymbol{F}_2 \, . \end{aligned} \tag{7.59}$$

As in the single constituent case, it would be possible to do this integral an-

alytically by ignoring the boundary prior $\psi(z, \boldsymbol{F})$. Converting back to the standard parametrisation, the final posterior would then be given by

$$P(\{z\} \mid \hat{\boldsymbol{F}}) \propto \mathcal{A}_2 M(z_1, z_2) \sum_k \sum_j w^k w^j c_2^{kj} \mathcal{N}(z_1, z_2 \mid \boldsymbol{\mu}_{\mathrm{z}}^{kj\hat{F}}, \underline{\boldsymbol{\Sigma}}_{\mathrm{zz}}^{kj\hat{F}}) \,. \tag{7.60}$$

With the boundary prior, the integral is no longer analytically tractable. As a result, we take the same approach as in the single constituent case and sample from the full, non-marginalised posterior. An additional complication here are the extra correlations factored into $M(z_1, z_2)$. As a result of this term, the posterior is no longer a Gaussian mixture and therefore does not permit the efficient sampling scheme described in section 7.1.4.

Instead, we can sample from the full posterior distribution ignoring the contribution of both the the boundary prior and the correlations, modifying the samples *post hoc* by rejection and reweighting to correct for these respectively. This set of samples is thus drawn from the simplified posterior $H(\boldsymbol{\psi})$, given by

$$\{z_1, z_2, \boldsymbol{F}_1, \boldsymbol{F}_2\} \sim H(\boldsymbol{\psi}) \propto \mathcal{A}_2 \sum_k \sum_j w^k w^j c_2^{kj} \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj\hat{F}}, \underline{\boldsymbol{\Lambda}}^{kj\hat{F}}) \,. \tag{7.61}$$

This simplified posterior is now a standard GMM, and can therefore be efficiently sampled as described in section 7.1.4. The neglected terms can now be corrected for separately.

Firstly, the boundary priors can be included by rejecting samples where the flux or the redshift is negative, as in section 7.2.1. The sorting condition could also be included by simply rejecting samples where it was not respected. However, this is unnecessarily wasteful of computation. Note that mixture component-jk is identical to component-kj under exchange of constituents. Every component is matched with a pair in this way. As a result, the posterior is exactly symmetric, meaning that samples with misordered redshifts can be corrected by simply swapping the order of their constituents.

The redshift correlation function can be corrected for using importance sampling by associating each sample with a weight $[1 + \xi(z_1, z_2)]$. All inferences done with these samples would then need to account for these weights. The risk with this importance sampling approach is that regions of parameter space where the correlation function is large could be poorly sampled when using the modified posterior. The effect of the correlation function would then be under-represented. However, chapter 6 found that including the redshift correlation function when sampling the posterior had little effect

on inferences. As a result, we expect any errors from the use of importance sampling here to be negligible.

Given a set of corrected samples of redshift and flux, the marginalisation can then be done in the same way as in section 7.2.1, by discarding the flux parts of the samples. The distribution of the remaining redshift samples will then be proportional to the marginalised posterior defined in equation 7.59, as desired.

7.2.4 Two-constituent evidence

The two-constituent evidence $\mathcal{E}^2$ is defined as the integral of the blended posterior over both sets of fluxes and redshifts, i.e.,

$$\mathcal{E}^2 = \iint P(\hat{\boldsymbol{F}} \mid \{\boldsymbol{F}\})\, P(\{z\}, \{\boldsymbol{F}\})\, \mathrm{d}\{z\}\, \mathrm{d}\{\boldsymbol{F}\}\,. \tag{7.62}$$

Inserting the definitions of each term from the full posterior given in equation 7.59, this expression becomes

$$\begin{aligned}\mathcal{E}^2 = \mathcal{A}_2 \iiiint & M(z_1, z_2)\psi(z_1, \boldsymbol{F}_1)\psi(z_2, \boldsymbol{F}_2)\times \\ & \sum_k \sum_j w^k w^j c_2^{kj} \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj\hat{F}}, \underline{\boldsymbol{\Lambda}}^{kj\hat{F}})\, \mathrm{d}z_1\, \mathrm{d}z_2\, \mathrm{d}\boldsymbol{F}_1\, \mathrm{d}\boldsymbol{F}_2\,.\end{aligned} \tag{7.63}$$

As before, we evaluate this integral numerically using Monte Carlo integration. To do this, we can reuse the samples drawn for the blended posterior inference from $H(\boldsymbol{\psi})$ defined in equation 7.61. Given a set of n_2 of these samples $\{z_1^i, z_2^i, \boldsymbol{F}_1^i, \boldsymbol{F}_2^i \mid i = 1 \ldots n_2\}$, we can define the weighted fraction

$$\mathcal{F}_2 = \sum_i \frac{[1 + \xi(z_1^i, z_2^i)]\; \pi(z_1^i, z_2^i)\psi(z_1^i, \boldsymbol{F}_1^i)\psi(z_2^i, \boldsymbol{F}_2^i)}{n_2}\,. \tag{7.64}$$

This is analogous to $\mathcal{F}_1$, the fraction of samples drawn from the non-marginalised single-constituent posterior defined in equation 7.28 that obey the boundary prior, but with the additional blending-related correlations. The simplified posterior $H(\boldsymbol{\psi})$ is not normalised to unity. However, the normalisation constant $\mathcal{V}_2$ is given by the integral over the full support of the distribution, giving

$$\begin{aligned}\mathcal{V}_2 &\equiv \int \mathcal{A}_2 \sum_k \sum_j w^k w^j c_2^{kj} \mathcal{N}(\boldsymbol{\psi} \mid \boldsymbol{\eta}^{kj\hat{F}}, \underline{\boldsymbol{\Lambda}}^{kj\hat{F}})\, \mathrm{d}\boldsymbol{\psi} \\ &= \mathcal{A}_2 \sum_k \sum_j w^k w^j c_2^{kj}\,.\end{aligned} \tag{7.65}$$

Thus, the two-constituent evidence can be estimated by importance sampling to be

$$\mathcal{E}^2 \approx \mathcal{V}_2\mathcal{F}_2 = \mathcal{A}_2\mathcal{F}_2 \sum_k \sum_j w^k w^j c_2^{kj} \,. \tag{7.66}$$

7.3 Tests on simulated sources

In order to test our method, we construct a two sets of simulated observations to train our model and compare predictions against. These two sets correspond to an LSST-like optical survey (Ivezić et al., 2019), and the same survey with additional Euclid-like infrared observations (Laureijs et al., 2011). The complementarity of LSST and Euclid has been investigated previously (e.g., Rhodes et al., 2017); additional filter bands will help to break colour-redshift degeneracies and therefore enable more accurate photometric redshifts.

Simulated observations are generated by redshifting a template, integrating over the relevant filter response curves, scaling the results to a given i-band magnitude, adding observational noise and imposing selection criteria. We use the set of templates assembled by Coe et al. (2006) containing eight templates. This is the default template set in the commonly used BPZ (Benítez, 2000) photometric redshift software.

We randomly sample true redshift, magnitude and template parameters for each source from a prior using `emcee` (Foreman-Mackey et al., 2013). The single-constituent joint redshift-magnitude-template prior is defined as follows. First we factorise into separate prior terms, i.e.,

$$P(z, m, t) = P(z \mid m)P(t \mid m)P(m) \,, \tag{7.67}$$

where t is an integer labelling each template and the redshift prior is assumed to be independent of template. The redshift and magnitude priors are then given by the LSST predictions in LSST Science Collaboration et al. (2009). The redshift prior, based on simulated high-redshift galaxy populations (Kitzbichler and White, 2007) is given by

$$P(z \mid m) = \frac{1}{2z_0(m)} \left(\frac{z}{z_0(m)}\right)^2 \exp\left(\frac{-z}{z_0(m)}\right) \,, \tag{7.68}$$

where

$$z_0(m) = 0.0417m - 0.744 \,, \tag{7.69}$$

and m refers to i-band magnitude. The corresponding i-band magnitude prior, fitted to data from the Canada-France-Hawaii Telescope Legacy Survey (CFHTLS; Hoekstra

et al., 2006), is then given by

$$P(m) \propto 10^{0.31(m-25)} \,. \tag{7.70}$$

We also use the template prior from Benítez (2000), given by

$$P\Big(t \,\Big|\, m\Big) = f_{\rm t} \exp\left(-k_{\rm t}[m - m_0]\right) \,, \tag{7.71}$$

where we set $m_0 = 20$ and the parameters $f_{\rm t}$ and $k_{\rm t}$, each dependent on the template type, are set to the values given in Benítez (2000).

Once the redshift, magnitude and template are sampled from this joint prior, the intrinsic fluxes are simulated by redshifting the template and integrating over filter response curves. For the optical survey, we use the six LSST filters u, g, r, i, z, Y (LSST Science Collaboration et al., 2009). We use the three Euclid filters Y, J, H (Racca et al., 2016) as additional infrared bands, giving a total of nine bands for the combined surveys.

Finally, we add magnitude-dependent observational noise to each band. For the optical bands, this is given by the predicted LSST noise model (LSST Science Collaboration et al., 2009). The 5σ depth of point sources in the Euclid Y, J and H bands is 24mag (Laureijs et al., 2011), the same depth as point sources in the LSST i-band (LSST Science Collaboration et al., 2009). We therefore approximate the observational noise in the Y, J and H bands by assuming that their signal-to-noise is equal to that of the i-band.

In order to simulate the flux of blended sources, we add the intrinsic fluxes of two simulated sources and add observational noise corresponding to the total blended flux. The two-constituent prior also needs to account for the blended-related terms described above. The redshift prior includes the sorting condition $\pi(z_1, z_2)$, though we assume no clustering, i.e., $\xi(z_1, z_2) = 0$, as it has a negligible effect at large separations when $z_1 \not\approx z_2$. We also impose a prior on the faintest i-band magnitude of either constituent such that it must be brighter than a 5σ detection. A cut like this is necessary since it only makes sense to consider a source blended when each constituent is sufficiently bright. If a constituent is too faint, it should instead be considered to be a contributor to the background flux, rather than that of the source itself.

Finally, we select sources by imposing an i-band magnitude cut of $m_i < 25$. This corresponds to the LSST gold sample (LSST Science Collaboration et al., 2009), a population of $\approx 4 \times 10^9$ high signal-to-noise galaxies. For each of the two sets of simulated sources, we randomly select 10000 single-constituent sources to act as a

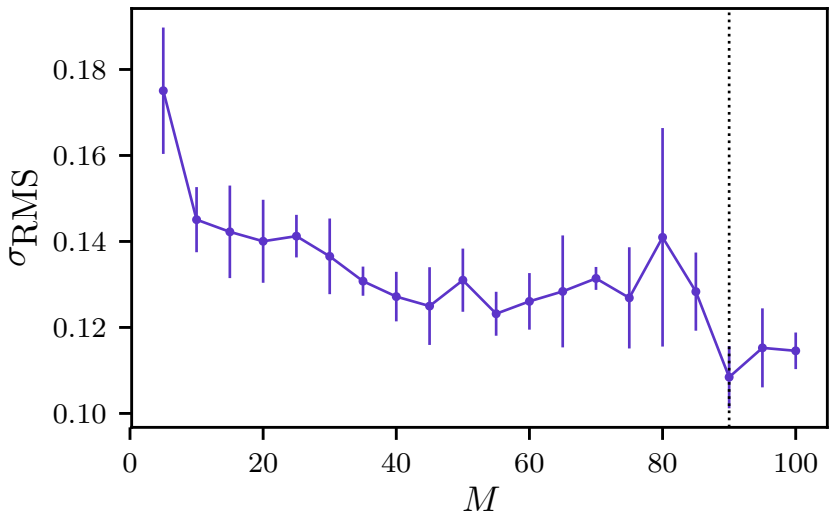

Figure 7.3: Results of the cross-validation for the LSST-like simulated data. The points show the RMS scatter averaged over the three folds, while the error bars show the error on the mean. We choose the number of components to be $N = 90$, minimising the average RMS scatter as indicated by the dotted black line.

training set, a further 10000 single-constituent sources for the unblended test set, and 10000 two-constituent sources for the blended test set.

Given the unblended training set, we use the procedure described in section 7.1.3 to set the number of mixture components N. Using 3-fold cross-validation, we test from $N = 5$ to $N = 100$ in multiples of 5, measuring the RMS scatter σ_{RMS} defined in equation 7.10 at each iteration. In order to evaluate this, we must define a way to calculate a point estimate z_{p} from a set of n_2 samples $\{z_{\mathrm{p},i} \mid i = 1 \ldots n_2\}$ drawn from the posterior defined in section 7.2.1. We therefore define this point estimate to be the mean of these samples, as this is equivalent to a Monte Carlo estimate of the expectation value of the redshift, i.e.,

$$z_{\mathrm{p}} \equiv \frac{1}{n_2} \sum_{i=1} z_{\mathrm{p},i} \approx \int P(z \mid \hat{\boldsymbol{F}})\, z \, \mathrm{d}z \,. \tag{7.72}$$

The results of this cross-validation are shown in Figure 7.3. We find the average RMS scatter across all folds $\overline{\sigma_{\mathrm{RMS}}}$ to be minimised when $N = 90$ with $\overline{\sigma_{\mathrm{RMS}}} = 0.108$. We therefore use a mixture comprised of 90 components to fit the entire training set for use throughout.

Examples of one-constituent posteriors inferred using samples from the distribution defined in section 7.2.1 and conditioned on the LSST-like data are shown in Figure 7.4. The four panels in this figure show the variety of shapes of posteriors that

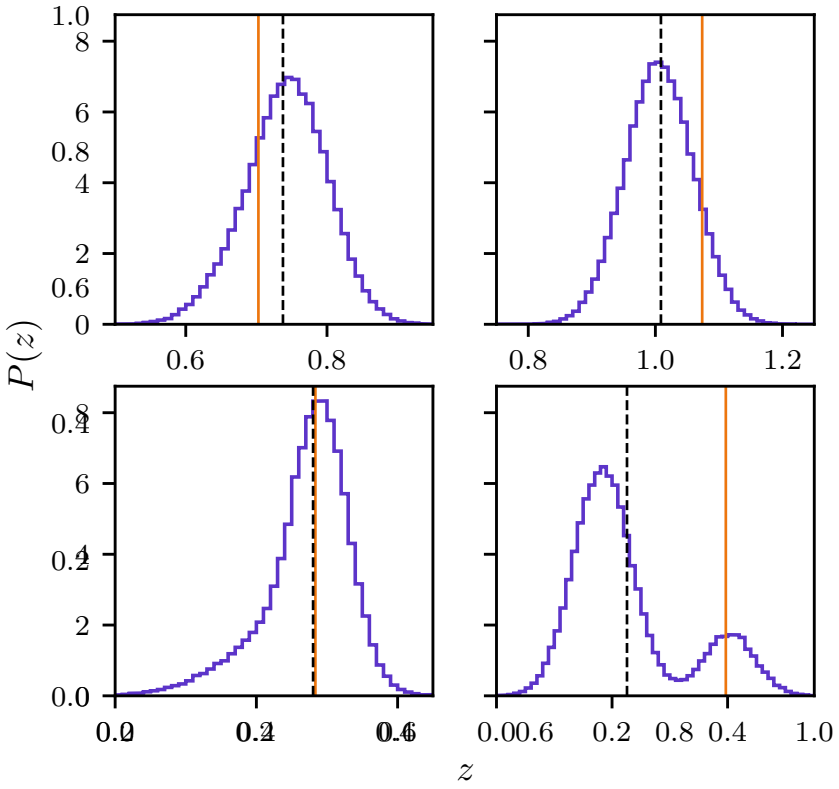

Figure 7.4: Plot showing four examples of single-constituent posteriors sampled using our method on the unblended LSST-like data. The black dashed lines indicate the sample means we use to define the point estimates z_{p}. The true redshifts are indicated by the solid orange lines.

can result from photometric redshifts and can be represented by the GMMs presented here.

The top two panels of Figure 7.4 shows examples of well constrained, accurate posteriors; their shapes are symmetric and close to that of a single Gaussian. However, the posterior shown in the bottom left panel is left-skewed. This long-tailed posterior is a common occurrence in the results of photometric redshift inference. Despite being very non-Gaussian, it can be represented by a mixture of components. Finally, the bottom right panel shows an example of a bimodal posterior that can be easily represented by a mixture of well separated components. While the true redshift is contained well within the lower peak of this posterior, the bimodality has pulled the mean redshift to between the two peaks. As a result, the point estimate is inaccurate, despite the true redshift lying at a point of significantly non-zero posterior density. This demonstrates the loss of information resulting from the compression of a full posterior distribution to a single point estimate.

Examples of two-constituent posteriors inferred using samples from the distribu-

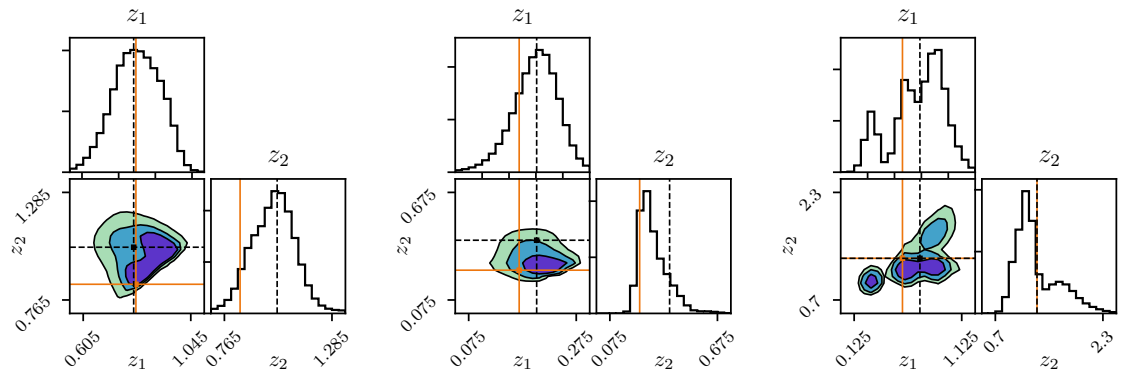

Figure 7.5: Plot showing three examples of two-constituent posteriors sampled using the GMM on the blended LSST-like data. The black dashed lines indicate the sample means we use to define the point estimates z_p. The true redshifts of each constituent are indicated by the orange lines.

tion defined in section 7.2.3 are shown in Figure 7.5. These samples are also drawn from posteriors conditioned on the LSST-like data.

The left panel of Figure 7.5 shows a well constrained posterior. One edge of the joint distribution lies along the $z_1 = z_2$ line. As a result, the effect of the sorting condition $\pi(z_1, z_2)$ can be seen clearly, sharply cutting the joint distribution. The centre panel shows a joint posterior that results in highly skewed marginal distributions. As before, the long tail of the z_2 marginal distribution pulls the mean redshift away from the peak. This demonstrates that, since point estimates are inevitably less informative than the full posterior distribution, the choice of how these point estimates are defined can significantly alter their accuracy. In this case, the accuracy of the point estimate would be increased by choosing z_2 to be the redshift where the posterior peaks, i.e., the maximum *a posteriori* (MAP) value. However, we found that MAP point estimates were less accurate over the whole sample on average. Finally, the right panel of Figure 7.5 shows an example of a highly multimodal posterior that can arise in the two-constituent case.

While less informative than the full posterior distributions, point estimates are still a common product of photometric redshift inference. A plot of these point estimates, defined as the mean of samples drawn from the posterior, against the true redshift for single-constituent data from the two simulated surveys is shown in Figure 7.6.

This figure shows that the method performs well in the single-constituent case, i.e., on the standard photometric redshift inference problem. The vast majority of sources have their redshifts recovered accurately; this can be seen by the significant density of points around the $z_\mathrm{p} = \hat{z}_\mathrm{s}$ line, demonstrated in the plot by the colour of the

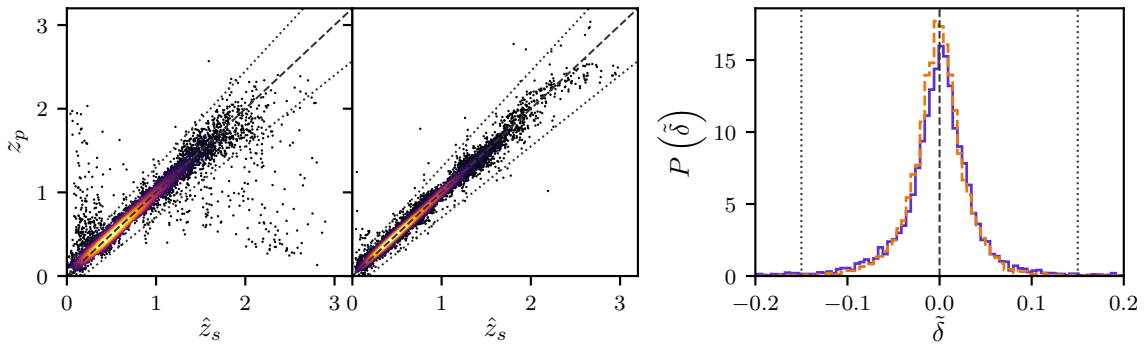

Figure 7.6: Plot showing the point-estimate results obtained from the GMM on the unblended simulated data. The left and right scatter plots show the point estimate results for the LSST-like and the combined LSST-Euclid-like surveys respectively. These plots show the benefit of additional bands and increased wavelength coverage from near-infrared data in reducing outliers. The dashed line denotes $z_\mathrm{p} = \hat{z}_\mathrm{s}$, and the dotted lines indicate our outlier definition where $|z_\mathrm{p} - \hat{z}_\mathrm{s}| \geq 0.15(1 + \hat{z}_\mathrm{s})$. Points are coloured according to their density on the scatter plots to illustrate overplotting. The right panel shows the distribution of the normalised error $\tilde{\delta}$, defined in equation 7.9. The solid purple line shows the results for the LSST-like survey, while the orange dashed line shows the results for the combined LSST-Euclid-like survey. The black dashed and dotted lines are defined as in the scatter plots.

points. Comparing the panels for the two simulations, the most significant difference is in the number of outliers, which is reduced in the simulations with additional infrared data. This can also be seen in the third panel, a histogram of the reduced error $\tilde{\delta}$ defined in equation 7.9. When zoomed around the majority of values at small errors, the difference between the histograms for the two sets of simulations is negligible.

This reduction of outliers is expected, as the additional filters can help to lift the colour-redshift degeneracies discussed in section 7.1. We define outliers to be sources where $|z_\mathrm{p} - \hat{z}_\mathrm{s}| \geq 0.15(1 + \hat{z}_\mathrm{s})$. This boundary is shown as a dotted line in Figure 7.6.

In order to quantify the accuracy of these point estimates, we can use several metrics. Firstly, we use the RMS scatter defined in equation 7.10. We find this scatter to be $\sigma_\mathrm{RMS} = 0.105$ for the LSST-like simulations, and $\sigma_\mathrm{RMS} = 0.038$ for the simulations with additional infrared data. While this difference is significant, it is primarily driven by the reduction of outliers by the infrared data.

In the LSST-like survey, 1.82% of sources are outliers. This is reduced to 0.10% in the combined LSST-Euclid-like simulations. These outliers have significant errors by definition, are therefore can have a significant effect on the measured RMS scatter. In order to identify these outliers as the most significant driver of the difference in accuracy between the two sets of simulations, we measure the RMS scatter while neglecting these sources, as in the photometric redshift accuracy tests of Hildebrandt et al. (2010).

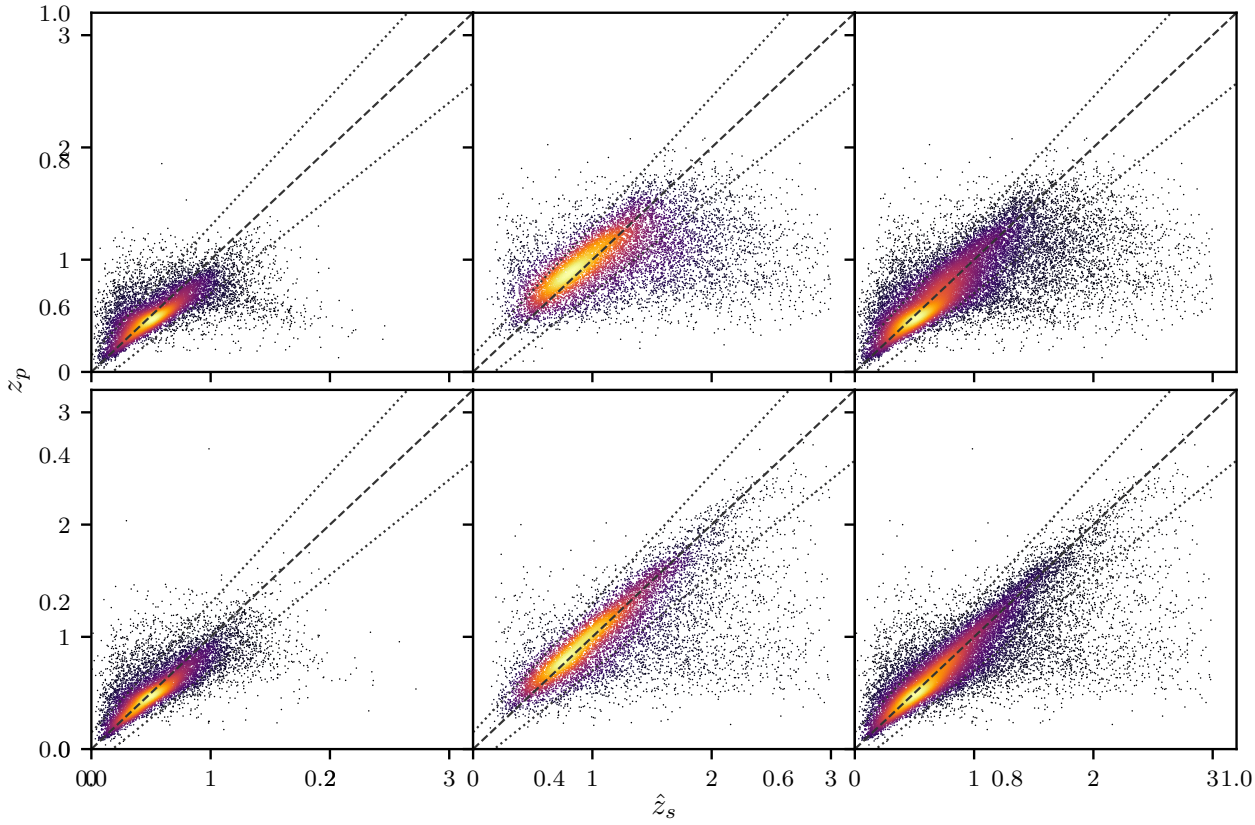

Figure 7.7: Plot showing the point-estimate results obtained from the GMM on the blended simulated data. The top row shows the results for the LSST-like survey, and the bottom row shows results for the combined LSST-Euclid-like survey. The left plots show $z_{\mathrm{p},1}$, the point estimate of the redshift for the lower-redshift constituent in each blended source. The centre plots show $z_{\mathrm{p},2}$, corresponding to the higher-redshift constituent in each blended source. The right plots combine both $z_{\mathrm{p},1}$ and $z_{\mathrm{p},2}$. The dashed lines denotes $z_{\mathrm{p}} = \hat{z}_{\mathrm{s}}$, and the dotted lines indicate our outlier definition where $|z_{\mathrm{p}} - \hat{z}_{\mathrm{s}}| \geq 0.15(1 + \hat{z}_{\mathrm{s}})$. Points are coloured according to their density on the scatter plots to illustrate overplotting.

When this is done, the RMS of the LSST-like simulations drops to $\sigma_{\mathrm{RMS}} = 0.036$, while the scatter of the simulations with additional Euclid-like data becomes $\sigma_{\mathrm{RMS}} = 0.031$. Since these values are now far closer and the latter change was less dramatic, we conclude that the biggest benefit afforded by the additional bands is the reduction of outliers.

We also evaluate the same metrics on point estimates of the redshifts of the blended simulated data. These point estimates are defined to be the mean of posterior samples, as in the single-constituent case. A plot of these point estimates for each set of simulated data is shown in Figure 7.7.

The blended redshift inference is a more challenging problem than standard photometric redshifts of unblended sources. However, while the scatter plots in Figure 7.7

are noisier than the single-constituent plots in Figure 7.6, many redshifts are still recovered accurately. This can be seen in the high density of points around $z_\mathrm{p} = \hat{z}_\mathrm{s}$, again demonstrated by their colour. This increase in noise over the single-constituent case is expected, as the same number of data-points per source are used here to constrain twice the number of parameters.

As in the single-constituent case, the addition of additional bands in the infrared reduces both the RMS scatter and the number of outliers. For the LSST-like survey, we find the scatter to be $\sigma_\mathrm{RMS} = 0.171$, while the combined LSST-Euclid-like survey has a scatter of $\sigma_\mathrm{RMS} = 0.145$. The outlier rate of the former survey is found to be 17.5%, while that of the latter is reduced to 12.4%.

As discussed in section 7.1, an important part of the results of photometric redshift inference are PDFs. Unlike simple point estimates, PDFs represent the full statistical knowledge of the redshift being inferred and are essential for rigorously propagating uncertainties. It is therefore also important that the quality of the resulting PDFs are assessed.

A conceptual problem with assessing the quality of PDFs is that there is no true PDF that they can be compared against. This is in contrast to point estimates where the spectroscopic redshift provides a known ground truth against which to compare. Instead, Wittman et al. (2016) introduce a frequentist method to test the widths of PDFs that relies on credible intervals (CIs).

The definition of CIs follows directly from that of posterior PDFs. For a given posterior $P(\psi \mid d)$ that is correctly normalised, the conditional probability that the parameter ψ will lie within an interval $[\psi_\mathrm{low}, \psi_\mathrm{high}]$ is given by the integral of the posterior over that interval, i.e.,

$$P(\psi_\mathrm{low} \leq \psi \leq \psi_\mathrm{high} \mid d) = \int_{\psi_\mathrm{low}}^{\psi_\mathrm{high}} P(\psi \mid d)\, \mathrm{d}\psi\,. \tag{7.73}$$

The CI corresponding to a particular percentage is then defined to be the interval over which equation 7.73 equals this percentage. In general, this interval will not be unique, since the integral over many different intervals can be the same. For this reason, the credible interval is often defined to be the highest posterior density (HPD) interval, the interval covering the shortest length in parameter space for a given integral. In general, this region does not need to be contiguous; the HPD region of multimodal posteriors will instead be made up of several subintervals.

A conceptually simple way to define this HPD region is to consider a horizontal line spanning the entirety of parameter space, drawn on a plot of the PDF. As this

line is moved downwards, it will begin to intersect the PDF. The regions between these intersections can then be integrated to give an area. The intervals contained within these intersections are the HPD region corresponding to this area. Since this area will monotonically increase as the line is moved downwards, this provides a way to define the HPD region for a given percentage CI.

An intuitive interpretation of these intervals is that, given many repetitions of the experiment and the subsequent construction of many such intervals of area α, the true parameter would be contained within a fraction α of these intervals. This notion is the interpretation of frequentist confidence intervals as coverage probabilities. However, while this interpretation is intuitive, it is not guaranteed by a Bayesian analysis. Instead, posteriors where this coverage probability property holds are said to be *calibrated*, with several methods having been proposed to calibrate posteriors (e.g. Syring and Martin, 2018; Sellentin and Starck, 2019).

The method introduced in Wittman et al. (2016) tests whether the posteriors resulting from a photometric redshift method are calibrated. If they are, we should expect that 50% of sources have their true redshift within their 50% CI. The equivalent statement can be made for all levels of CI, generalising this to a continuous test. The test may therefore give an indication of the performance of the method, and such a test has been widely adopted in the photometric redshift literature (e.g., Leistedt and Hogg, 2017; Gomes et al., 2017; Duncan et al., 2018; Meshcheryakov et al., 2018; Amaro et al., 2018; Rodríguez-Muñoz et al., 2019).

By definition, if the true redshift of a source lies within its 50% CI, it will also lie within all CIs corresponding to larger percentages, as the 50% CI will be a subset of these. It is therefore sufficient to measure only the threshold CI that just contains the true redshift. This will have one of the interval edges at the true redshift. This region can therefore be measured by drawing the horizontal line detailed above so that it intersects the posterior at the true redshift. The area c corresponding to this interval is measured for each galaxy in the sample being tested. The cumulative distribution function (CDF) of these areas CDF(c) can then be calculated. Wittman et al. (2016) note that for calibrated posteriors, the plot of this CDF against areas should be diagonal, i.e., CDF(c) = c. The deviation away from this line therefore measures how overconfident or underconfident the PDFs are.

A plot of this test for the LSST-like simulated data is shown in Figure 7.8. This figure shows that both the one- and two-constituent posteriors are approximately calibrated and their CIs can therefore be interpreted in a frequentist manner.

Finally, Figure 7.9 shows the relative probability for the blended and unblended

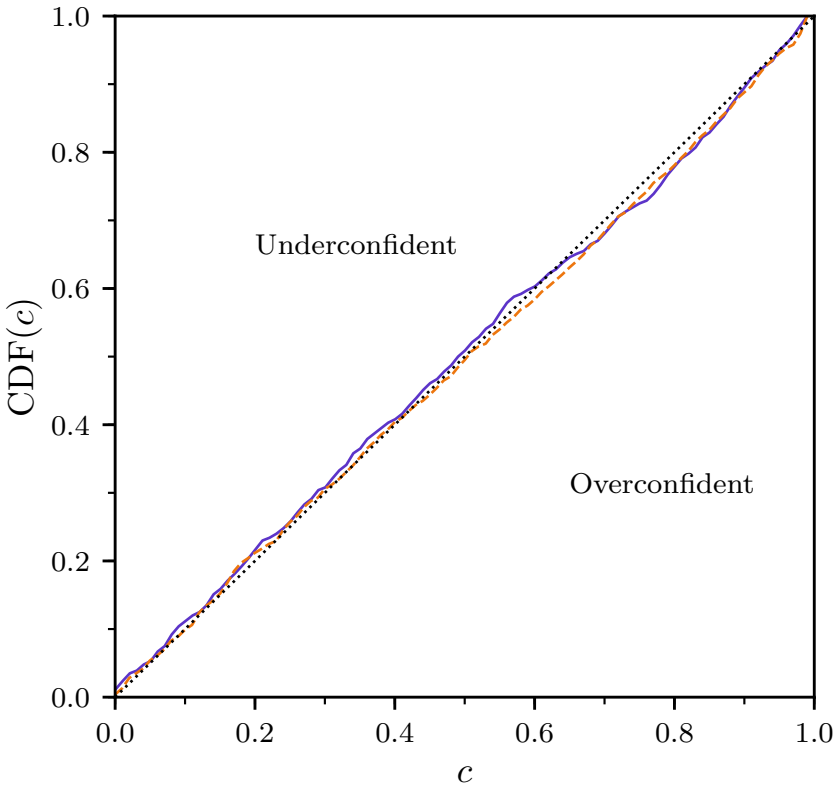

Figure 7.8: Plot showing the results of the posterior width test performed on posteriors obtained from our method on LSST-like simulated data. The solid purple line shows the results for the single-constituent posteriors, and the dashed orange line shows the results for the two-constituent posteriors. The black dotted line indicates the result where posteriors are calibrated, while lines that go above and below this indicate posteriors that are wider and narrower than calibrated posteriors respectively.

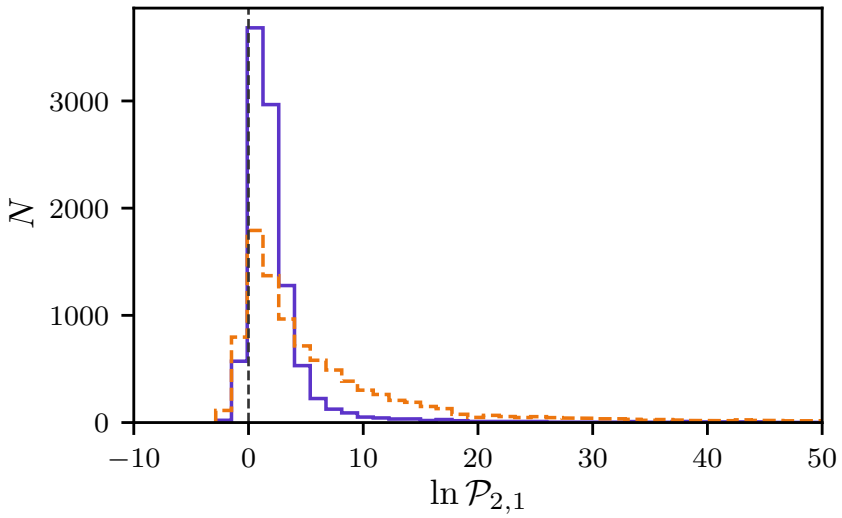

Figure 7.9: Histograms of the log of the relative probabilities for the blended and unblended models obtained using Bayesian model comparison on the simulated blended data. The solid purple histogram shows the result for the LSST-like survey, while the dashed orange histogram shows the result for the combined LSST-Euclid-like survey. The black dashed line indicates no preference for either the unblended or blended model. Larger values of $\mathcal{P}_{2,1}$ favour the blended model more.

models $\mathcal{P}_{2,1}$ calculated for the blended data of both simulated surveys. This quantity is calculated using the evidences derived in sections 7.2.2 and 7.2.4 using equation 6.43. We assume a ratio of model priors of unity, i.e., we do not *a priori* favour either the one- or two-constituent models. A blended source is then favoured when $\ln \mathcal{P}_{2,1} > 1$. We find that the LSST-like survey identifies 92.4% of blended sources, while the survey with additional infrared data identifies 89.3%.

7.4 GAMA blended sources catalogue

In addition to the simulated observations presented in section 7.3, we also test our method against real observations. To do this, we use data from the Galaxy And Mass Assembly (GAMA) survey (Baldry et al., 2017), a spectroscopic survey of $> 150\,000$ sources. Alongside this spectroscopy, these sources were also imaged in optical wavelengths by the Sloan Digital Sky Survey (SDSS) (Stoughton et al., 2002) and in infrared wavelengths by the VISTA Kilo-degree Infrared Galaxy (VIKING) Survey (Edge et al., 2013). Hill et al. (2011) used this imaging data to create self-consistent, aperture-matched photometry in nine bands $u, g, r, i, z, Y, J, H, K$ for all sources within the GAMA survey. As a result, these sources have both high-quality photometry and accurate spectroscopic redshifts for training and testing our photometric redshift method.

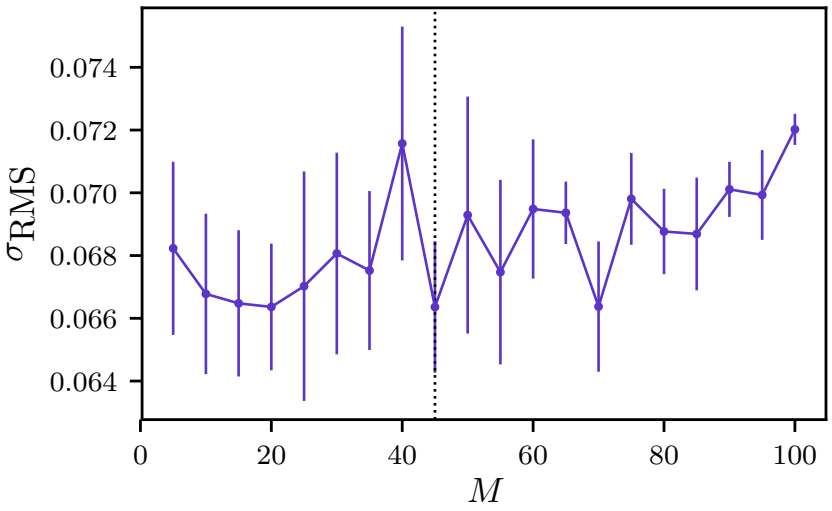

Figure 7.10: Results of the cross-validation for the GAMA blended sources catalogue data. The points show the RMS scatter averaged over the three folds, while the error bars show the error on the mean. We choose the number of components to be $N = 45$, minimising the average RMS scatter as indicated by the dotted black line.

Holwerda et al. (2015) used this data to spectroscopically identify blended sources in order to search for strong-lens candidates. The resulting GAMA blended sources catalogue contains blended photometry for 280 sources, alongside the spectroscopic redshift of each constituent. We therefore use this catalogue to test the performance of our method on real observations of blended sources. To accompany this, we also randomly select two sets of 10000 unblended sources for a training and test set.

As for the simulated observations, we use 3-fold cross-validation to find the number of mixture components N that minimises $\overline{\sigma_{\mathrm{RMS}}}$ the RMS scatter averaged over all folds. The results of this are shown in Figure 7.10. We find the minimum scatter when the number of mixture components is $N = 45$, giving $\overline{\sigma_{\mathrm{RMS}}} = 0.066$. We therefore continue with a GMM of 45 components fitted to the 10000 unblended training sources.

We then compute point estimates of the single-constituent redshifts by averaging samples drawn from the posterior as before. A plot of this is shown in Figure 7.11. We find the RMS scatter to be $\sigma_{\mathrm{RMS}} = 0.067$, with 3.6% of sources being outliers.

A scatter plot of the two-constituent point estimates is shown in Figure 7.12. As in the simulated case, the blended results are noisier than the single-constituent case. We find the RMS scatter to be $\sigma_{\mathrm{RMS}} = 0.091$, and 10.8% of sources to be outliers.

Examples of single-constituent posteriors are shown in Figure 7.13. Like the single-constituent posteriors conditioned on the simulated data, these distributions

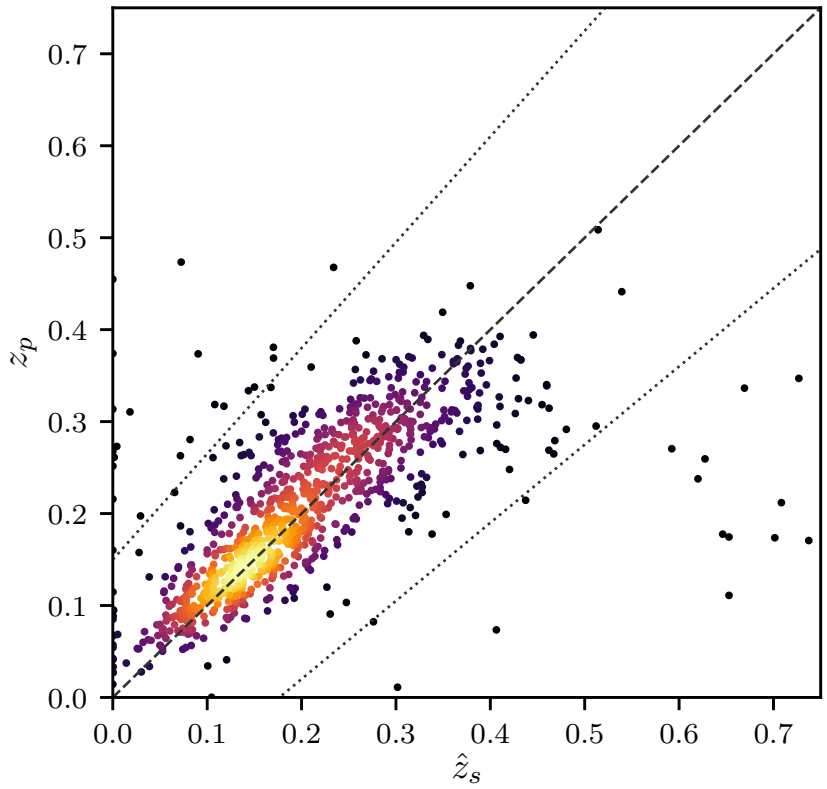

Figure 7.11: Plot showing the point-estimate results obtained from the GMM on the unblended GAMA data. The dashed line denotes $z_{\rm p} = \hat{z}_{\rm s}$, and the dotted lines indicate our outlier definition where $|z_{\rm p} - \hat{z}_{\rm s}| \geq 0.15(1 + \hat{z}_{\rm s})$. Points are coloured according to their density on the scatter plots to illustrate overplotting.

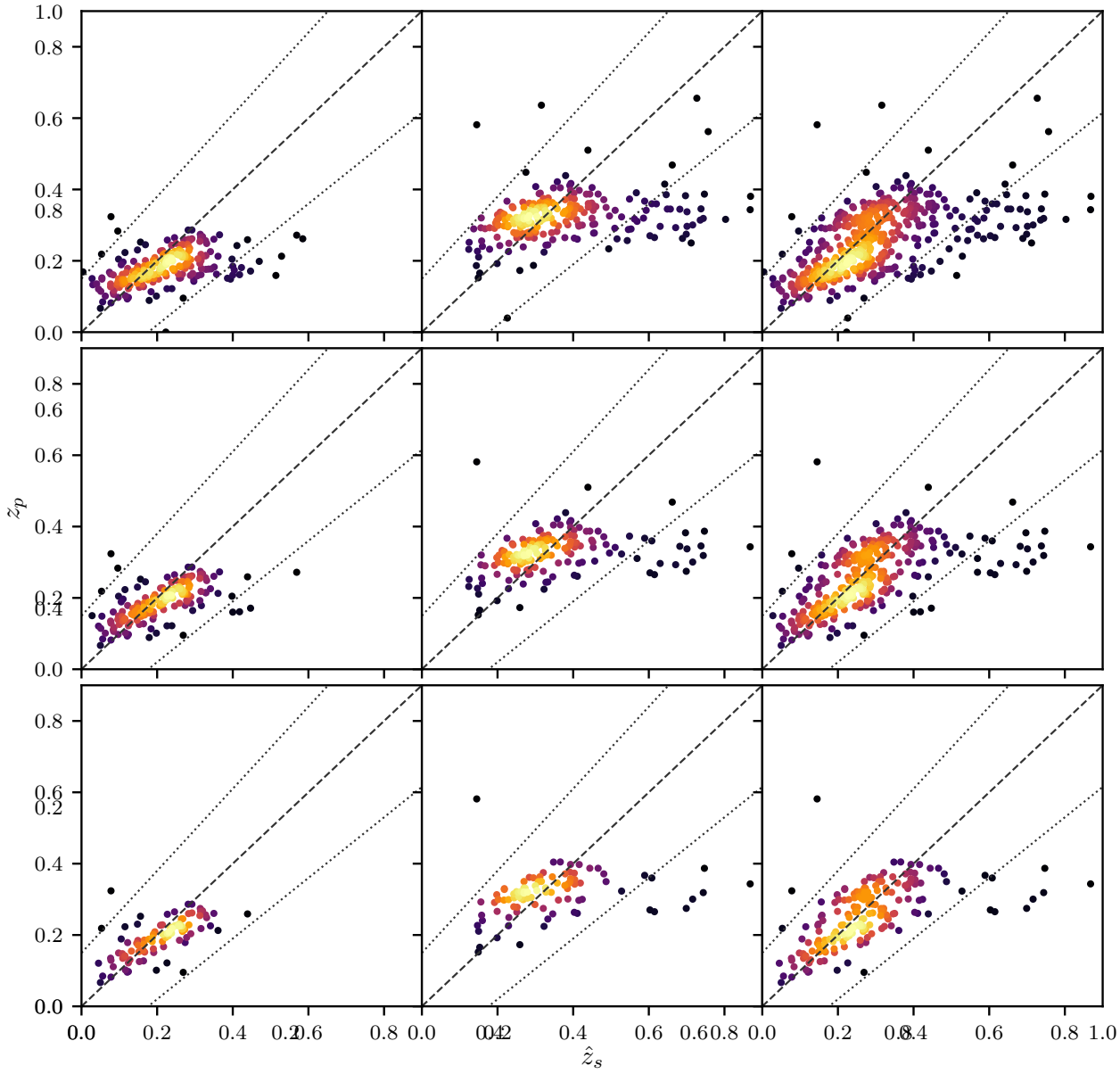

Figure 7.12: Plot showing the point-estimate results obtained from the GMM on the data from the GAMA blended sources catalogue, with various density ratio thresholds. The left column shows $z_{\mathrm{p},1}$, the point estimate of the redshift for the lower-redshift constituent in each blended source. The centre column shows $z_{\mathrm{p},2}$, corresponding to the higher-redshift constituent in each blended source. The right column combines both $z_{\mathrm{p},1}$ and $z_{\mathrm{p},2}$. The top row shows the results for the full sample, while the centre and bottom rows have sources with expected density ratios less than 0.45 and 0.8 removed respectively. where the expected density ratio is defined in equations 7.76 and 7.77. Imposing this density ratio threshold removes sources that are least well-represented in the training set, and so we would expect the results to improve as the threshold is increased. As indicated in the text, the summary statistics improve as expected by making these cuts. This can also be seen visually in this figure by comparing the lower two rows with the full sample in the top row. The dashed lines denotes $z_{\mathrm{p}} = \hat{z}_{\mathrm{s}}$, and the dotted lines indicate our outlier definition where $|z_{\mathrm{p}} - \hat{z}_{\mathrm{s}}| \geq 0.15(1 + \hat{z}_{\mathrm{s}})$. Points are coloured according to their density on the scatter plots to illustrate overplotting.

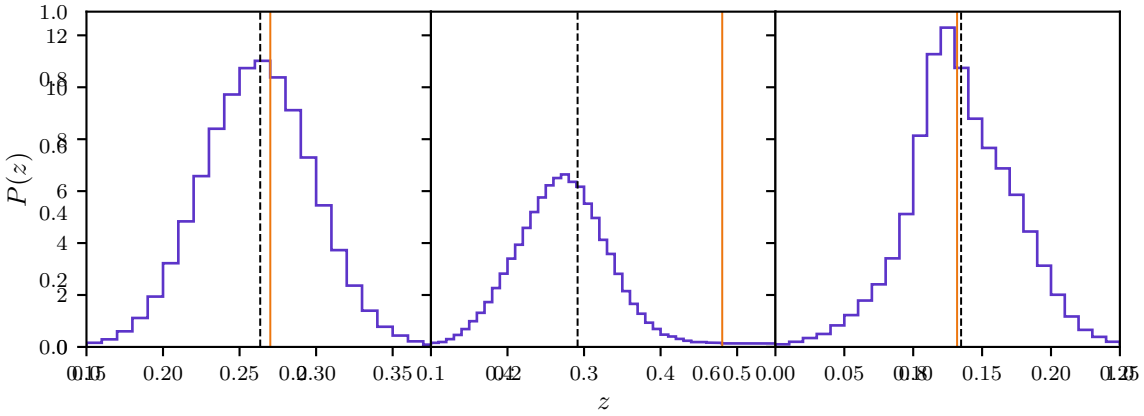

Figure 7.13: Plot showing three examples of single-constituent posteriors sampled using the GMM on the unblended GAMA data. The black dashed lines indicate the sample means we use to define the point estimates z_{p}. The true redshifts are indicated by the orange lines.

show a variety of shapes. However, the posteriors for the GAMA data are significantly less multimodal. This is likely because the GAMA sources are, on average, lower redshift than the simulated sources. The main cause of the bimodality in the simulated case is the colour-redshift degeneracy described in section 7.1, which low- and high-redshift sources to be confused. However, high redshifts are *a priori* very unlikely here, as they do not appear in the training set. As a result, these higher redshift peaks are significantly disfavoured.

The same lack of multimodality is also exhibited in the blended posteriors conditioned on the GAMA data. Examples of these are shown in Figure 7.14. These posteriors show a variety of non-Gaussian shapes as in the simulated case, with many of the marginal redshift distributions displaying long tails. The joint distribution in the left panel of Figure 7.14 also shows the hard cut resulting from the sorting condition $\pi(z_1, z_2)$, as the left panel of Figure 7.5 does.

Figure 7.15 shows the plot testing the posterior widths for both the one- and two-constituent posteriors. As in the simulated case, the one-constituent posteriors are very close to being calibrated. However, the CDF for the two-constituent posteriors lies significantly below the diagonal, suggesting that the posteriors are overconfident, i.e., they are too narrow. As discussed above, while it is not guaranteed that Bayesian CIs provide frequentist coverage probabilities, this suggests that there are features on the flux-redshift relation of the blended constituents that are not captured by the model trained on the unblended training data.

This interpretation is supported by Figure 7.16 which shows the inferred relative probability of sources from the blended sources catalogue being blended and unblended

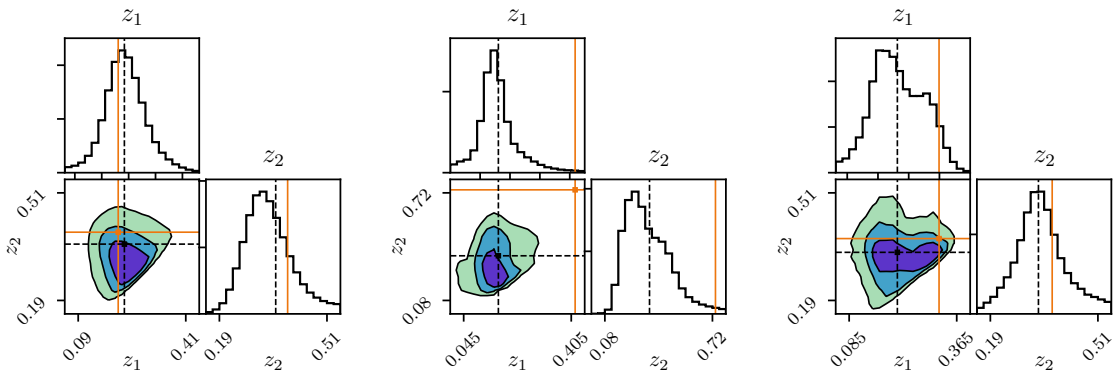

Figure 7.14: Plot showing three examples of two-constituent posteriors sampled using the GMM on data from the GAMA blended sources catalogue. The black dashed lines indicate the sample means we use to define the point estimates z_p. The true redshifts of each constituent are indicated by the orange lines.

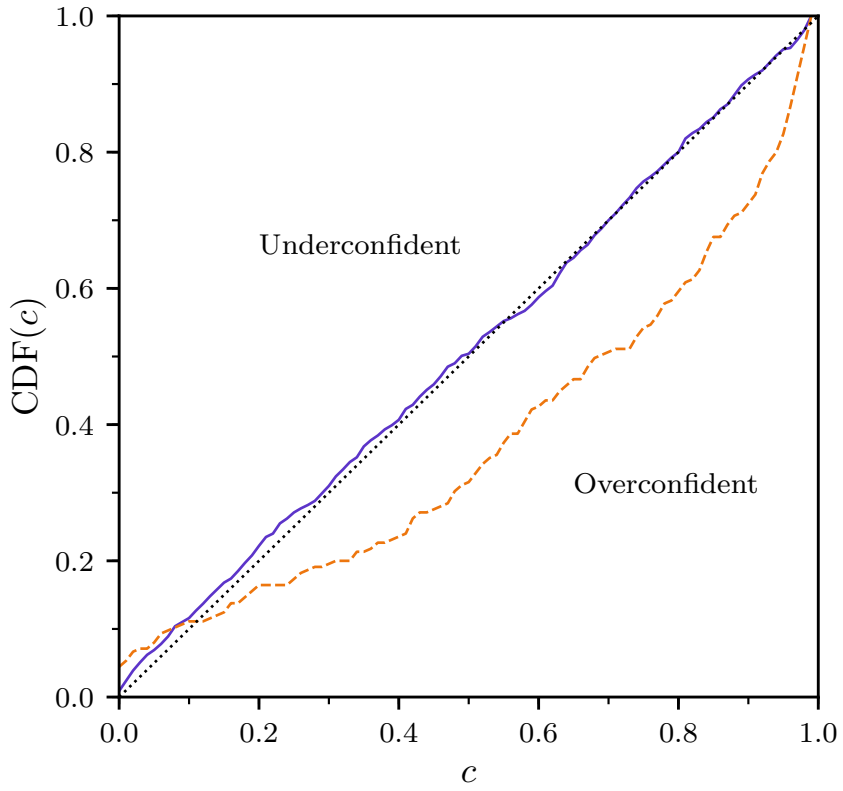

Figure 7.15: Plot showing the results of the posterior width test performed on posteriors obtained from our method on GAMA data. The solid purple line shows the results for the single-constituent posteriors, and the dashed orange line shows the results for the two-constituent posteriors. The black dotted line indicates the result where posteriors are calibrated, while lines that go above and below this indicate posteriors that are wider and narrower than calibrated posteriors respectively.

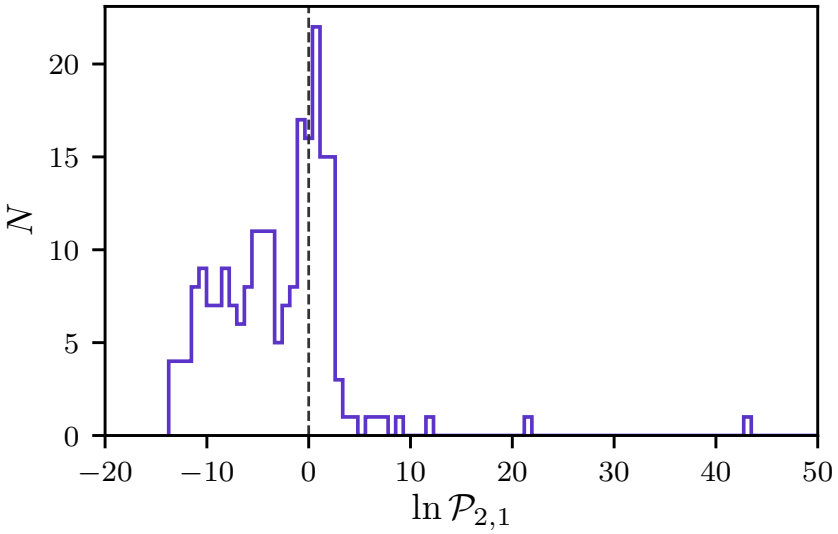

Figure 7.16: Histogram of the log of the relative probabilities for the blended and unblended models obtained using Bayesian model comparison on the blended GAMA data. The black dashed line indicates no preference for either the unblended or blended model. Larger values of $\mathcal{P}_{2,1}$ favour the blended model more.

$\mathcal{P}_{2,1}$. Here, only 33.4% of blended sources are correctly identified as blended by having $\mathcal{P}_{2,1} > 1$. While the redshifts are reasonably well-recovered, the Bayesian model selection will disfavour a more complicated model when the improvement in the fit is insufficient. As above, this suggests a difference between the blended and unblended constituents.

We can test for a difference between the blended and unblended constituents by incrementally removing sources where this difference is greatest and checking whether this leads to an improvement in the summary statistics. We therefore require a quantity to probe the representativeness of a given vector of fluxes. For this, we consider the density ratio

$$\mathcal{R}(\boldsymbol{F}) = \frac{P_{\rm test}(\boldsymbol{F})}{P_{\rm train}(\boldsymbol{F})}\,, \tag{7.74}$$

where $P_{\rm train}(\boldsymbol{F})$ is the density of fluxes is the training set, $P_{\rm train}(\boldsymbol{F})$ is the density of fluxes is the test set and $\boldsymbol{F}$ is the flux vector at which both of these densities are evaluated.

In order to estimate this ratio, we use the nearest-neighbour method of Kremer et al. (2015). The method first considers the training set, and measures the hypervolume that contains the $n_{\rm nei}$ nearest neighbours of a flux $\boldsymbol{F}$. The number of test-set samples $n_{\rm test}(\boldsymbol{F})$ within that hypervolume centred on $\boldsymbol{F}$ is then counted. The estimate

for the density ratio is then given as the ratio of these counts, i.e.,

$$\mathcal{R}(\boldsymbol{F}) \approx \frac{n_{\rm nei}}{n_{\rm test}(\boldsymbol{F})}\,. \tag{7.75}$$

This nearest-neighbour method for estimating the density ratio was first presented in Lima et al. (2008), and was used to estimate the redshift distribution of a photometric galaxy sample by weighting spectroscopic galaxies. However, the accuracy of this method depends on $n_{\rm nei}$, the number of neighbours considered. If $n_{\rm nei}$ is too large, the density ratio is estimated over too large a volume, while an estimate where $n_{\rm nei}$ is too small will be dominated by statistical errors. To this end, Kremer et al. (2015) present a model-selection method based on cross-validation to optimise $n_{\rm nei}$.

As discussed throughout this chapter, a complication of blended sources is that the flux of each constituent is not observed independently, only the blended combination. As a result, the destiny ratio must be evaluated using constituent fluxes sampled from the marginal posterior $P(\boldsymbol{F}_n|\hat{\boldsymbol{F}})$, where $\boldsymbol{F}_n$ is the flux of constituent n. As described in section 7.2.3, this can be accomplished by sampling from the simplified posterior defined in equation 7.61, and rejecting samples that do not obey the boundary prior. The marginalisation over all redshifts and the flux of the other constituent can than be done by simply ignoring these elements of the sampled vectors.

Given a set of $n_{\rm F}$ flux samples $\{\boldsymbol{F}_n^i \mid i = i \ldots n_{\rm F}\}$ from constituent n, we evaluate the density ratio $\mathcal{R}(\boldsymbol{F})$ for each sample and average the result to give the expectation value

$$\mathrm{E}[\mathcal{R}(\boldsymbol{F})] \equiv \int \mathcal{R}(\boldsymbol{F}_n) P(\boldsymbol{F}_n|\hat{\boldsymbol{F}})\,\mathrm{d}\boldsymbol{F}_n \approx \frac{1}{n_{\rm F}} \sum_i \mathcal{R}(\boldsymbol{F}_n^i)\,. \tag{7.76}$$

This expectation value is the quantity we use to estimate the representativeness of blended constituents. This allows us to test for differences between the blended and unblended constituents. To do this, we keep sources in our sample only if the expectation of the density ratio for both of their constituents is over a threshold value $\mathcal{R}_{\rm th}$, i.e., sources that obey

$$\frac{\mathrm{E}[\mathcal{R}(\boldsymbol{F}_n^i)]}{\max(\mathrm{E}[\mathcal{R}(\boldsymbol{F})])} \geq \mathcal{R}_{\rm th}, \quad n \in \{1, 2\}\,, \tag{7.77}$$

where we have normalised the expectation values by $\max(\mathrm{E}[\mathcal{R}(\boldsymbol{F})])$, the maximum expectation value over both constituents of all sources.

Figure 7.17 shows the change in summary statistics as the threshold ratio is increased. As expected, the RMS scatter and number of outliers are both reduced as this ratio is increased, at the expense of more sources being removed from the

sample. This effect can also be seen in the lower two rows of Figure 7.12, where the effects of two different threshold values on the point estimates are compared with the unmodified results. When the threshold is set at $\mathcal{R}_{\rm th} = 0.45$ as in the centre row, the RMS scatter has been reduced to $\sigma_{\rm RMS} = 0.078$, while the percentage of sources that are outliers has reduced to 5.97%. At this level, 70.7% of sources remain in the sample. By increasing the threshold to $\mathcal{R}_{\rm th} = 0.8$ as in the bottom row, the RMS scatter and percentage of outliers decrease to $\sigma_{\rm RMS} = 0.077$ and 4.34% respectively. These are modest improvements over the less strict threshold, but come at the cost of leaving only 40.9% of sources remaining in the sample.

These results demonstrate the importance of representative training sets. Differences between the training and test sets, often referred to as covariate shift, are a general problem for machine learning-based methods that obtain all of their information from the training set. A possible cause of differences here is that surveys select sources based on a magnitude cut, imparting selection effects on the sample. Since blended sources will be selected based on their total blended flux, blended constituents can be fainter than those that are unblended. The simulated sources presented in section 7.3 are selected in this way and so contain this effect. However, the intrinsic properties of galaxies vary with magnitude, meaning that the test set could contain faint constituents that have no corresponding examples in the test set. Selection effects imparted by the selection criteria of sources in the blended sources catalogue, such as certain redshift differences being easier to select spectroscopically, are also not accounted for here.

One solution to this problem is to improve the training set so that it is more representative. By including sources in the training set fainter than the magnitude limit of the test set, the model can learn the faint-end flux-redshift relation. The selection effects of blended sources could also be learned directly by training using a blended training set as described in section 7.1.1. However, as detailed above, assembling a representative blended training set in practice could be difficult. For the tests presented here, the GAMA blended sources catalogue contains far too few sources to be amenable to fitting in this way.

7.5 Conclusions

Future galaxy surveys will observe to unprecedented depths in order to drive their increases in precision of cosmological constraints. However, these improvements to constraints on cosmological parameters will be accompanied by several new complica-

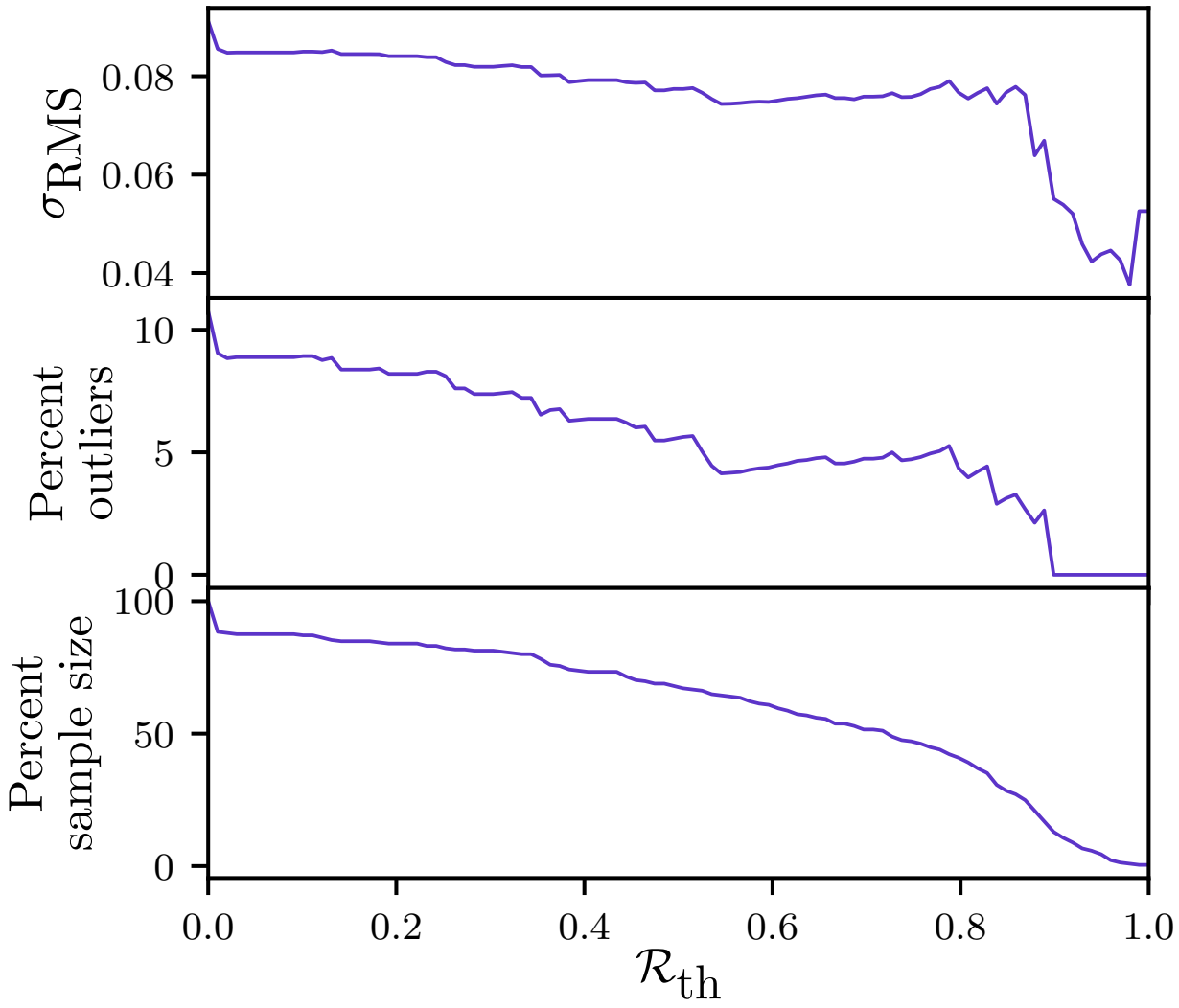

Figure 7.17: Plot showing the change in summary statistics for the GAMA blended sources as the density ratio threshold $\mathcal{R}_{\rm th}$ is increased. The top panel shows the RMS scatter $\sigma_{\rm RMS}$. The centre panel shows the percentage of sources that are outliers, defined as $|z_{\rm p} - \hat{z}_{\rm s}| \geq 0.15(1 + \hat{z}_{\rm s})$. The bottom panel shows the percentage of sources remaining from the original sample after the threshold has been applied.

tions to the analysis. The increased number density of sources will increase both the number of sources that are blended and the total number of sources observed.

This chapter presents a photometric redshift method for blended sources based on Gaussian mixture models. Using these models, our method learns the flux-redshift distribution from a set of unblended training galaxies. This choice of model permits the derivation of posteriors that can be sampled efficiently, allowing the method to scale to large samples. By using Bayesian model selection techniques, this method can also infer the number of constituents within a blended sources efficiently.

This work extends previous uses of GMMs in photometric redshift applications (Bovy et al., 2012) to the case of blended sources. It also extends the template-based method to infer the redshifts of blended sources directly from their blended photometry first introduced in chapter 6. The method described therein relies on nested sampling for inference and so will not scale to the large sample sizes of future galaxy surveys such as LSST (Ivezić et al., 2019). The method presented in this chapter is significantly faster, making it suitable for these upcoming surveys. Many modern methods of photometric redshifts are machine learning-based, as training these methods on a representative training set can allow them to achieve very high accuracy and avoid the problems associated with small template sets. This chapter extends the blended photometric redshift method of chapter 6 to this data-driven approach.

The accuracy of all machine learning-based photometric redshift methods is dependent of the training set. Using training sets that are unrepresentative could result in redshift inferences that are biased and posterior distributions that are too narrow. In cases where unblended galaxies are not representative of individual components in a blended source, potentially as a result of selection effects, our method can generalise to learn the blended flux-redshift relation directly from blended training data. While this naturally accounts for differences between blended and unblended galaxies, it also increases the size of the required training set.

The method presented here represents a different approach to analysing blended sources than is currently used. Rather than separating blended observations into separate constituents, we infer the redshifts jointly for all constituents. As a result, our method naturally captures uncertainties and correlations which can be difficult to estimate for deblending-based analyses. This approach could be extended to other quantities of interest for cosmological analysis such as galaxy shapes by constructing forward models of source images. By doing this, correlations associated with blending can be propagated fully throughout the rest of the analysis, providing the best understanding of uncertainties on cosmological constraints.

As with chapter 6, while the discussions in this chapter focus on galaxy-galaxy blending, the method presented here could be applied to blends with other types of objects such as stars and quasars. In fact, Bovy et al. (2012) use their GMM-based model both to obtain photometric redshifts for quasars and as a probabilistic classifier to distinguish between quasars and stars. To apply our model to these problems, we would require a representative training set of each type of object on which separate models would be trained. The model comparison procedure described above could then be used to between models representing different combinations of object types.

Chapter 8

Bayesian Hierarchical Model for Blended Redshift Distributions

Chapters 6 and 7 present methods that generalise photometric redshift methods to the case of blended sources. The aim of both of these methods is to infer the posterior distribution of the redshifts of individual sources, where the number of constituents in those sources is unknown. However, as discussed in section 2.3, for many cosmological applications of galaxy surveys, it is not only the redshifts of individual sources that are required, but also the distribution of redshifts for the entire population of observed sources.

This distinction between photometric redshifts for individual sources and populations is discussed in more detail in chapter 4. One of the methods detailed therein is the use of a Bayesian Hierarchical Model (BHM), where the prior of a Bayesian model is itself parametrised by parameters that are inferred and included within the final posterior distribution. This allows observations of samples from a population to be used to make inferences about the population itself. These methods can therefore be applied to the problem of photometrically inferring the redshift distribution of a population using flux observations of individual sources. BHMs are introduced more generally in section 3.1.8.

This chapter described a method that builds on the Gaussian Mixture Model (GMM) method described in chapter 7, extending it to the task of inferring redshift distributions for populations of possibly blended sources. The model infers separate redshift distributions for unblended sources in addition to the lower and higher redshift constituents in blended sources. Section 8.1 describes our hierarchical GMM approach to inferring these blended redshift distributions. Section 8.2 discusses a slightly modified model that describes these distributions as discrete histograms rather than

continuously. Section 8.3 shows the results of tests of these methods on simulated data. Finally, section 8.4 discusses some possible further work which could extend these methods.

8.1 Hierarchical Gaussian mixture model

The method presented in chapter 7 for inferring redshifts for each source independently assumes that the joint flux-redshift distribution is the same for unblended galaxies and for each galaxy in a two-component blended source. However, that need not be the case. The selection effects on each galaxy in a blended source differ from those on unblended galaxies by virtue of them having been selected on the total blended flux, rather than their individual fluxes alone. The effect of this is that the constituent galaxies present in blended sources can be fainter than unblended sources. This will inevitably have an effect on the redshift distributions of these galaxies too, since higher redshift galaxies will tend to be fainter.

In order to model this, we construct a Bayesian hierarchical model over the three distinct populations, treating the priors on unblended sources and each component in blended sources separately. Like the GMM approach described in chapter 7, the relationship between fluxes and redshifts is learned from a training set of unblended galaxies with spectroscopically obtained redshifts. We therefore begin by fitting a GMM to represent the joint flux-redshift distribution of the training set, given by

$$P(z, \boldsymbol{F}) = \sum_k w_k^{\mathrm{tr}} \, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) \,, \tag{8.1}$$

where we label the weight vector $\boldsymbol{w}^{\mathrm{tr}}$ to indicate that it corresponds to the training set. As in chapter 7, this GMM is fitted using the extreme deconvolution method (Bovy et al., 2011); this is discussed in more detail in section 7.1.1.

Next, to describe the unblended sources and each component in blended sources, our model consists of three GMM priors. We assume that the mean vectors and covariance matrices of these mixtures are fixed at their training set values $\{\boldsymbol{\mu}\}$ and $\{\underline{\boldsymbol{\Sigma}}\}$ as in equation 8.1, and are therefore same between each prior; future work that would relax this assumption is described in section 8.4.3. Each GMM prior then differs only through its weights. These weights are labelled $\boldsymbol{w}^1$, $\boldsymbol{w}^\alpha$ and $\boldsymbol{w}^\beta$, corresponding to the unblended sources, the lower-redshift component of the blended sources, and the higher-redshift component of the blended sources respectively. Our aim is therefore to derive a posterior for the these three vectors of weights conditioned on the set of test-set

fluxes $\{\hat{\boldsymbol{F}}\}$. Given one of these weight vectors, the corresponding redshift distribution is given by a GMM as in equation 8.1, marginalised over the flux distribution, i.e.,

$$P(z \mid \boldsymbol{w}^*) \equiv \int P(z, \boldsymbol{F} \mid \boldsymbol{w}^1)\, \mathrm{d}\boldsymbol{F} = \sum_k w_k^*\, \mathcal{N}(z, \boldsymbol{F} \mid \boldsymbol{\mu}_{\mathrm{z}}^k, \underline{\boldsymbol{\Sigma}}_{\mathrm{zz}}^k)\,, \tag{8.2}$$

where $\boldsymbol{w}^* \in [\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta]$ is the weight vector corresponding to the particular distribution, and $\boldsymbol{\mu}_{\mathrm{z}}^k$ and $\underline{\boldsymbol{\Sigma}}_{\mathrm{zz}}^k$ are the redshift parts of the k^{th}-component mean vector and covariance matrix respectively.

To develop our desired posterior, we first apply Bayes rule to give

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto P(\{\hat{\boldsymbol{F}}\} \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta)\,, \tag{8.3}$$

where we suppress the conditioning on the means and covariances for conciseness. We then make the assumption that the fluxes are i.i.d., allowing us to rewrite the likelihood as a product over the data, giving

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \prod_i P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta)\,. \tag{8.4}$$

The weights $\boldsymbol{w}^1$, $\boldsymbol{w}^\alpha$ and $\boldsymbol{w}^\beta$, along with the fixed means and covariances, parametrise the flux-redshift distribution for single-component sources and for each component in two-component blended sources respectively.

8.1.1 Developing the posterior through model averaging

Since we are fitting the weights to the test-set fluxes, we do not know *a priori* whether or not each source is blended. To account for this, we use Bayesian model averaging. As described in section 3.1.6, this technique introduces a latent parameter $\mathcal{M}_i$ representing the choice of model which can be marginalised over. The posterior assuming a particular model is then conditioned on it, i.e., $P(\boldsymbol{\theta} \mid \{\boldsymbol{d}\}, \mathcal{M}_i)$, where $\boldsymbol{\theta}$ are the model parameters and $\{\boldsymbol{d}\}$ is the dataset.

Like marginalising over any other parameter, model averaging incorporates the epistemic uncertainty into the final posterior and thus the parameter inferences. The usual prescription, not applicable here, is to compute the model averaged posterior, given by

$$\begin{aligned} P(\boldsymbol{\theta} \mid \{\boldsymbol{d}\}) &= \sum_i P(\boldsymbol{\theta}, \mathcal{M}_i \mid \{\boldsymbol{d}\}) \\ &= \sum_i P(\theta \mid \{\boldsymbol{d}\}, \mathcal{M}_i) P(\mathcal{M}_i \mid \{\boldsymbol{d}\})\,, \end{aligned} \tag{8.5}$$

where the first term is the posterior under model $\mathcal{M}_i$ and the second is the probability for that model. As described in section 3.1.6, this model probability is given by

$$P(\mathcal{M}_i \mid \{\boldsymbol{d}\}) = \frac{P(\{\boldsymbol{d}\} \mid \mathcal{M}_i)P(\mathcal{M}_i)}{\sum_i P(\{\boldsymbol{d}\} \mid \mathcal{M}_i)P(\mathcal{M}_i)}\,. \tag{8.6}$$

In the case of model averaging, it is assumed that the true model is contained within the set $\{\mathcal{M}_i\} = \{\mathcal{M}_1, \mathcal{M}_2 \ldots \mathcal{M}_N\}$ that is marginalised over.

In this way, the uncertainty over the choice of model can be included in the final posterior. However, note that equation 8.5 marginalises over the model at the level of the full dataset. It is assumed that only a single unknown model is responsible for the generation of the entire dataset $\{\boldsymbol{d}\}$. Since that model is unknown, the uncertainty from the choice of model should be incorporated into the posterior by marginalising.

This assumption is not suitable for our problem, where the choice of model is the number of constituents in a source. Since each source may be either a single source or blended, there is no single model that describes the entire dataset of N sources. As a result, in contrast to standard model averaging, we assume the model can vary at the level of individual samples. To do this, we rewrite the likelihood by marginalising over N latent model parameters $\mathcal{C}_i$, representing the number of constituents for sample i to give

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) = \sum_{j=1}^{2} P(\hat{\boldsymbol{F}}_i, \mathcal{C}_i = j \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \tag{8.7}$$

Inserting this into the posterior, it becomes

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \prod_i \left[\sum_{j=1}^{2} P(\hat{\boldsymbol{F}}_i, \mathcal{C}_i = j \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \right]. \tag{8.8}$$

Further developing this joint $\hat{\boldsymbol{F}}_i, \mathcal{C}_i$ likelihood presents a problem. This likelihood could be factorised in the form $P(\hat{\boldsymbol{F}}_i, \mathcal{C}_i \mid \ldots) = P(\hat{\boldsymbol{F}}_i \mid \mathcal{C}_i \ldots)P(\mathcal{C}_i \mid \ldots)$. The flux likelihood $P(\hat{\boldsymbol{F}}_i \mid \mathcal{C}_i \ldots)$ is then conditioned on the number of constituents $\mathcal{C}_i$; this is necessary as this number controls which forward model and weight vectors are used to evaluate the likelihood. However, conditioning in this way means that the $P(\mathcal{C}_i \mid \ldots)$ distribution does not depend on the observed flux for each source. As a result, developing the posterior in this way is unsuitable for inferring the blending probability of each source from its observed flux, and would instead require this probability be specified *a priori* using external data such as the morphology.

In this situation where the desired posterior can be written most readily in terms of several conditional distributions, Gibbs sampling can be used to sample from

the full joint posterior distribution. Here, it is possible to specify the problem in terms of the joint posterior on both the weights and the number of components, i.e., $P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \{\mathcal{C}\} \mid \{\hat{\boldsymbol{F}}\})$, where $\{\mathcal{C}\}$ is the full set of N parameters. Since writing the conditional distributions of each of these is tractable, this distribution is amenable to Gibbs sampling. While this vastly increases the total number of parameters, each blended probability is a simple discrete parameter $\mathcal{C}_i \in \{1, 2\}$ which could be easily sampled without rejection.

While Gibbs sampling would allow samples to be drawn from the desired posterior, doing so would necessitate the creation of a custom Gibbs sampler. Due to time constraints, this was not possible. Instead, we simplify the posterior so that it can be sampled using standard sampling packages such as Stan (Carpenter et al., 2017). To do this, we first note that, assuming a uniform prior over the weights, we can write

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \hat{\boldsymbol{F}}_i)\,, \tag{8.9}$$

since the normalisations $P(\{\hat{\boldsymbol{F}}\}) = \prod_i P(\hat{\boldsymbol{F}}_i)$ are equal due to the i.i.d. assumption. The term of the right-hand side is then the sample-i posterior corresponding to a single source i. This posterior can then be marginalised over the number of constituents $\mathcal{C}_i$ as described above, giving

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i \left[\sum_j P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = j \mid \hat{\boldsymbol{F}}_i)\right]. \tag{8.10}$$

Finally, rewriting the term inside the square brackets using product rule, this becomes

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i^N \left[\sum_{j=1}^2 P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \hat{\boldsymbol{F}}_i, \mathcal{C}_i = j) P(\mathcal{C}_i = j \mid \hat{\boldsymbol{F}}_i)\right]. \tag{8.11}$$

The result of this manipulation is that the probability of the number of the number of constituents $P(\mathcal{C}_i = j \mid \hat{\boldsymbol{F}}_i)$ is now conditioned on the observed flux as desired. However, it is no longer conditioned on the weight vectors $\boldsymbol{w}^1$, $\boldsymbol{w}^\alpha$ and $\boldsymbol{w}^\beta$. Instead, we can calculate this probability using the model comparison techniques described in chapter 7 by making the simplifying assumption that this probability is conditioned on the previously fitted training set weights $\boldsymbol{w}^{\text{tr}}$, i.e., $P(\mathcal{C}_i \mid \{\hat{\boldsymbol{F}}\}) \equiv P(\mathcal{C}_i \mid \boldsymbol{w}^{\text{tr}}, \{\hat{\boldsymbol{F}}\})$. As a result, the blending probability is constant w.r.t. the parameters $\boldsymbol{w}^1$, $\boldsymbol{w}^\alpha$ and $\boldsymbol{w}^\beta$. As with the means and covariances, we suppress the conditioning on $\boldsymbol{w}^{\text{tr}}$ in our notation throughout.

By making this assumption of the blending probability remaining fixed, the pos-

terior distribution for the weight vectors can be written in terms of one joint distribution. This distribution is then amenable to being sampled using the Hamiltonian Monte Carlo (HMC) sampler Stan (Carpenter et al., 2017). This significantly simplifies the implementation of the method. Stan also provides an implementation of black box variational inference (VI, Kucukelbir et al., 2015) as described in section 3.2.4. VI is an approximate inference method that is significantly faster than HMC at the expense of no longer being exact. Both of these inference methods are tested in section 8.3. We now continue to develop the posterior using this approximation.

8.1.2 Approximating the blending probability as fixed

The term inside the square brackets in equation 8.11 now resembles the standard model averaging case shown in equation 8.5, where the first term is the sample-i posterior under a model of $\mathcal{C}_i$ components and the second is the probability for this number of components. We can then develop this individual posterior by applying Bayes rule to give

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i^N \left[\sum_{j=1}^2 P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = j) P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = j) P(\mathcal{C}_i = j \mid \hat{\boldsymbol{F}}_i) \right] . \tag{8.12}$$

We now explicitly expand the sum over the number of components to give

$$\begin{aligned} P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i^N \Big[& P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 1) P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = 1) P(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i) \\ & + P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = 2) P(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i) \Big] . \end{aligned} \tag{8.13}$$

We can now make some simplifications since each set of weights corresponds only to a single model. As a result, these weights only appear in the likelihood of their corresponding model, i.e.,

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 1) \equiv P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) \tag{8.14}$$

and

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) \equiv P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) . \tag{8.15}$$

We can also make similar simplifications for the weight priors. Firstly, we assume that the weight vectors are independent under both models, i.e.,

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = j) \equiv P(\boldsymbol{w}^1 \mid \mathcal{C}_i = j) P(\boldsymbol{w}^\alpha \mid \mathcal{C}_i = j) P(\boldsymbol{w}^\beta \mid \mathcal{C}_i = j)\,, \tag{8.16}$$

for $j \in \{1, 2\}$. We now set the priors for the weight vectors such that each weight vector can only have an influence under its respective model, i.e.,

$$P(\boldsymbol{w}^1 \mid \mathcal{C}_i = 2) = P(\boldsymbol{w}^\alpha \mid \mathcal{C}_i = 1) = P(\boldsymbol{w}^\beta \mid \mathcal{C}_i = 1) = 1\,. \tag{8.17}$$

Thus, the joint priors are given by

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = 1) \equiv P(\boldsymbol{w}^1 \mid \mathcal{C}_i = 1) \tag{8.18}$$

and

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \mathcal{C}_i = 2) \equiv P(\boldsymbol{w}^\alpha \mid \mathcal{C}_i = 2) P(\boldsymbol{w}^\beta \mid \mathcal{C}_i = 2)\,. \tag{8.19}$$

The final expression for the posterior then becomes

$$\begin{aligned} &P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \\ &\quad \prod_i^N \Big[P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) P(\boldsymbol{w}^1 \mid \mathcal{C}_i = 1) P(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i) \\ &\qquad + P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) P(\boldsymbol{w}^\alpha \mid \mathcal{C}_i = 2) P(\boldsymbol{w}^\beta \mid \mathcal{C}_i = 2) P(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i) \Big]\,. \end{aligned} \tag{8.20}$$

8.1.3 Specifying the priors

We now specify the priors on the weights. These weights have several constraints imposed on them; each weight w_i should be within the interval $0 \leq w_i \leq 1$, and the weights should be normalised such that $\sum_i w_i = 1$. A D-dimensional vector of weights therefore lies on the unit $(D-1)$-simplex, as demonstrated by Figure 8.1. As a result of this, the weight prior should have support over this simplex.

The natural choice for this is the Dirichlet distribution, which can be thought of as a distribution over discrete distributions. Our weights would then therefore distributed like

$$\begin{aligned} P(\boldsymbol{w}^1 \mid \mathcal{C}_i = 1) &= \mathrm{Dir}(\boldsymbol{w}^1 \mid \boldsymbol{a}^1) \\ P(\boldsymbol{w}^\alpha \mid \mathcal{C}_i = 2) &= \mathrm{Dir}(\boldsymbol{w}^\alpha \mid \boldsymbol{a}^\alpha) \\ P(\boldsymbol{w}^\beta \mid \mathcal{C}_i = 2) &= \mathrm{Dir}(\boldsymbol{w}^\beta \mid \boldsymbol{a}^\beta) \end{aligned} \tag{8.21}$$

where the vectors $\boldsymbol{a}^1$, $\boldsymbol{a}^\alpha$, $\boldsymbol{a}^\beta$ are hyperparameters, and the D-dimensional PDF is given

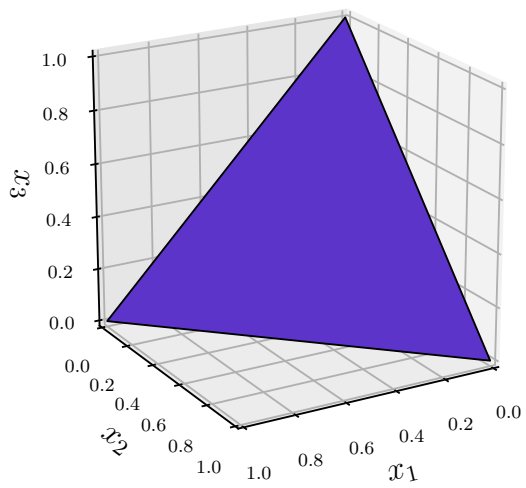

Figure 8.1: Surface plot showing a unit 2-simplex embedded within three-dimensional space. This simplex is the support of the three-dimensional Dirichlet distribution.

by

$$\mathrm{Dir}(\boldsymbol{x} \mid \boldsymbol{a}) = \frac{\Gamma(\sum_{i=1}^{D} \alpha_i)}{\prod_{i=1}^{D} \Gamma(\alpha_i)} \prod_{i=1}^{D} x_i^{\alpha_i - 1} \,. \tag{8.22}$$

where $\Gamma(\dots)$ is the gamma function. When all of the elements of $\boldsymbol{a}$ are equal, no component is favoured *a priori*. This parameter is then given as a scalar a known as a *concentration parameter*, and the corresponding distribution $\mathrm{Dir}(\boldsymbol{w} \mid \alpha)$ is said to be *symmetric*.

When a symmetric D-dimensional Dirichlet distribution is parametrised with a concentration parameter $a = 1$, the resulting distribution is uniform over the unit $(D-1)$-simplex, i.e., all sets of weights that obey the normalisation and nonnegative conditions are equally likely. When the concentration parameter $a < 1$, this favours samples that are more concentrated towards extreme values of the weights, meaning that the corresponding discrete distribution is peaked. When the concentration parameter $a > 1$, samples where the weights are closer in value are preferred, meaning that the corresponding discrete distribution is smooth. This behaviour is demonstrated in the ternary plots shown in Figure 8.2, which display samples from a three-dimensional symmetric Dirichlet distribution with several concentration parameters.

While the fixed blending-probability approximation made above relies on the assumption of a uniform prior over the weights, naively using this prior would violate the constraints on the weights described above. However, we note that a simple way to transform a uniformly distributed set of weights $\boldsymbol{w}$ into a set of weights $\boldsymbol{w}'$ that obey these constraints is to define $w_i' = {}^{|w_i|}/_{\sum_j |w_j|}$. Thus, by replacing every instance of the

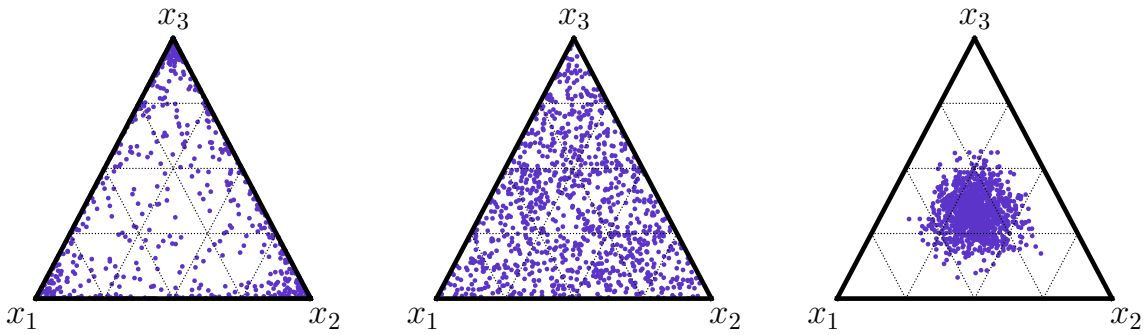

Figure 8.2: Ternary plots of samples drawn from a symmetric three-dimensional Dirichlet distribution $\mathrm{Dir}(\boldsymbol{x} \mid \alpha)$ with varying α. The left plot shows $\alpha = 0.25$, where samples are pushed towards having elements with extreme values. The centre plot shows $\alpha = 1$, indicating a uniform distribution over the simplex. The right plot shows $\alpha = 10$, where samples having closer elements is preferred. Each vertex of the triangle corresponds to a dimension i of the vector $\boldsymbol{x}$, where points at the vertex indicate $x_i = 1$, points at the opposite edge indicate $x_i = 0$ and the lines parallel to the opposite edge indicate constant x_i.

weights in the model with this redefinition, the model could be written to incorporate a uniform prior over the weights. The resulting set of weights $\boldsymbol{w}'$ will be uniformly distributed over the simplex of weights that obey the nonnegative and normalisation conditions required of the weights. As a result, this model with a uniform prior and the corresponding redefined weights is equivalent to the original model where the weights have a symmetric Dirichlet distribution with a concentration parameter $\alpha = 1$.

Avoiding the redefinition of the weights increases the clarity of presentation of the model. In addition, implementations of the Dirichlet distribution within HMC software such as Stan (Carpenter et al., 2017) are designed to use transformations to make the resulting posterior geometry easier to explore (e.g., Betancourt, 2012). We therefore assume the weights have symmetric Dirichlet priors with $\alpha = 1$ throughout this chapter.

8.1.4 One-constituent likelihood

We now derive the single-constituent likelihood $P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1)$. We start by introducing and marginalising over latent parameters for the true redshift and fluxes, giving

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) = \int\int P(\hat{\boldsymbol{F}}_i, \boldsymbol{F}_i, z_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1)\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i \tag{8.23}$$

Applying product rule, this becomes

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) &= \int\int P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i, z_i, \boldsymbol{w}^1, \mathcal{C}_i = 1) P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1)\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i \\ &= \int\int P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i, \mathcal{C}_i = 1) P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}^1)\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i \end{aligned} \tag{8.24}$$

where we have removed unnecessary conditioning. The first term of the integrand is the flux likelihood, a multivariate Gaussian centred on the observed fluxes, i.e.,

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i, \mathcal{C}_i = 1) = \mathcal{N}(\boldsymbol{F}_i \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i})\,, \tag{8.25}$$

where $\underline{\boldsymbol{\Sigma}}^{\hat{F}_i}$ is the covariance matrix of the flux observations. The second term of the integrand is the Gaussian mixture prior including the boundary prior as described in chapter 7, given by

$$P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) = \mathcal{A}_1 \psi(z, \boldsymbol{F}_i) \sum_k w_k^1\, \mathcal{N}(\boldsymbol{F}_i, z_i \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}^k)\,. \tag{8.26}$$

where $\psi(z, \boldsymbol{F})$ is the boundary prior defined in equation 7.7, and $\mathcal{A}_1$ is the prior normalisation defined in equation 7.26. The product of these two terms can be written as another Gaussian mixture as in equation 7.20. Thus, the single-component likelihood becomes

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) = \mathcal{A}_1 \sum_k w_k^1\, \mathcal{N}(\boldsymbol{\mu}_f^k \mid \hat{F}_i, \underline{\boldsymbol{\Sigma}}_{ff}^k + \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \times \\ \int\int \psi(z, \boldsymbol{F}_i) \tilde{\mathcal{N}}(\boldsymbol{F}_i, z_i \mid \boldsymbol{\eta}^{ik}, \underline{\boldsymbol{\Lambda}}^{ik})\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i \end{aligned} \tag{8.27}$$

where the combined parameters $\boldsymbol{\eta}^{ik}$ and $\underline{\boldsymbol{\Lambda}}^{ik}$ are defined as they are in equations 7.21 and 7.22. We label the integral over flux and redshift by I_{ik}, i.e.,

$$I_{ik} \equiv \int\int \psi(z, \boldsymbol{F}_i) \tilde{\mathcal{N}}(\boldsymbol{F}_i, z_i \mid \boldsymbol{\eta}^{ik}, \underline{\boldsymbol{\Lambda}}^{ik})\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i \tag{8.28}$$

As in chapter 7, this integral is computed numerically to account for the boundary prior $\psi(z, \boldsymbol{F})$. Without this term, the Gaussian density is normalised to unity, i.e., $\int\int \tilde{\mathcal{N}}(\boldsymbol{F}_i, z_i \mid \boldsymbol{\eta}^{ik}, \underline{\boldsymbol{\Lambda}}^{ik})\, \mathrm{d}\boldsymbol{F}_i\, \mathrm{d}z_i = 1$. Thus, given a set of N samples $\{z_j, \boldsymbol{F}_j \mid j = 1 \ldots N\}$ sampled from this density, the integral I_{ik} can be approximated as the fraction $\mathcal{F}_{ik}$ of these samples within the boundary, i.e.,

$$I_{ik} \approx \frac{\sum_{j=1}^N \psi(z_j, \boldsymbol{F}_j)}{N}\,. \tag{8.29}$$

We also relabel the normalisation term in equation 8.27 arising from the product of Gaussian densities to be

$$\gamma_{ik} = \mathcal{N}(\boldsymbol{\mu}_f^k \mid \hat{F}_i, \underline{\boldsymbol{\Sigma}}_{ff}^k + \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \,. \tag{8.30}$$

Thus, inserting these into equation 8.27, the single-constituent likelihood becomes

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) &\equiv \mathcal{A}_1 \sum_k w_k^1 \, \gamma_{ik} I_{ik} \\ &\equiv \mathcal{L}_i^1(\boldsymbol{w}^1) \end{aligned} \tag{8.31}$$

where we use the shortened notation $\mathcal{L}_i^1(\boldsymbol{w}^1)$ to refer to the one-constituent likelihood for source i.

8.1.5 Two-constituent likelihood

We now compute the two-constituent likelihood using the same approach of introducing and marginalising out latent parameters. With the extra constituent, there are now twice as many parameters to marginalise, so this likelihood is given by

$$\begin{aligned} &P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \\ &\qquad \iiiint P(\hat{\boldsymbol{F}}_i, \boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) \, \mathrm{d}\boldsymbol{F}_i^\alpha \, \mathrm{d}z_i^\alpha \, \mathrm{d}\boldsymbol{F}_i^\beta \, \mathrm{d}z_i^\beta \end{aligned} \tag{8.32}$$

As before, we separate the integrand by applying product rule to obtain

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \iiiint & P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta, \mathcal{C}_i = 2) \times \\ & P(\boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) \, \mathrm{d}\boldsymbol{F}_i^\alpha \, \mathrm{d}z_i^\alpha \, \mathrm{d}\boldsymbol{F}_i^\beta \, \mathrm{d}z_i^\beta \end{aligned} \tag{8.33}$$

Similarly to the single-component case, there are two probability densities here. The first is a multivariate Gaussian likelihood, given by

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta, \mathcal{C}_i = 2) = \mathcal{N}(\boldsymbol{F}_i^\alpha + \boldsymbol{F}_i^\beta \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \,. \tag{8.34}$$

where $\underline{\boldsymbol{\Sigma}}^{\hat{F}_i}$ is the covariance matrix of the flux observations. The second term of the integrand in equation 8.33 is the joint prior. As in the previous chapters we separate the joint prior by constituent, where the correlations between these components are assumed to be described completely by the redshift correlation function $\xi(z_i^\alpha, z_i^\beta)$ and the sorting condition $\pi(z_i^\alpha, z_i^\beta)$. The joint prior then contains two copies of the Gaussian

mixture priors, each also including the boundary prior described in chapter 7. This joint prior is therefore given by

$$P(\boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \mathcal{A}_2 \psi(z_i^\alpha, \boldsymbol{F}_i^\alpha)\psi(z_i^\beta, \boldsymbol{F}_i^\beta)[1 + \xi(z_i^\alpha, z_i^\beta)] \times \\ \pi(z_i^\alpha, z_i^\beta) \sum_k \sum_j w_k^\alpha w_j^\beta \mathcal{N}(\boldsymbol{F}_i^\alpha, z_i^\alpha \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}^k)\mathcal{N}(\boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{\mu}_j, \underline{\boldsymbol{\Sigma}}^j)\,, \tag{8.35}$$

where $\mathcal{A}_2$ is the prior normalisation defined in equation 7.48. Inserting these terms, the two-constituent likelihood becomes

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \mathcal{A}_2 \sum_k \sum_j w_k^\alpha w_j^\beta \iiiint \psi(z_i^\alpha, \boldsymbol{F}_i^\alpha)\psi(z_i^\beta, \boldsymbol{F}_i^\beta) \times \\ \pi(z_i^\alpha, z_i^\beta)[1 + \xi(z_i^\alpha, z_i^\beta)]\mathcal{N}(\boldsymbol{F}_i^\alpha + \boldsymbol{F}_i^\beta \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \times \\ \mathcal{N}(\boldsymbol{F}_i^\alpha, z_i^\alpha \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}^k)\mathcal{N}(\boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{\mu}_j, \underline{\boldsymbol{\Sigma}}^j)\, \mathrm{d}\boldsymbol{F}_i^\alpha\, \mathrm{d}z_i^\alpha\, \mathrm{d}\boldsymbol{F}_i^\beta\, \mathrm{d}z_i^\beta \tag{8.36}$$

As in equation 7.55 during the derivation of the two-component posterior in chapter 7, this product of three densities here can be written as another Gaussian mixture given by

$$P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \mathcal{A}_2 \sum_k \sum_j w_j^\alpha w_k^\beta \gamma_{ijk} \iiiint [1 + \xi(z_i^\alpha, z_i^\beta)]\pi(z_i^\alpha, z_i^\beta) \times \\ \psi(z_i^\alpha, \boldsymbol{F}_i^\alpha)\psi(z_i^\beta, \boldsymbol{F}_i^\beta)\tilde{\mathcal{N}}(\boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{\eta}^{ijk}, \underline{\boldsymbol{\Lambda}}^{ijk})\, \mathrm{d}\boldsymbol{F}_i^\alpha\, \mathrm{d}z_i^\alpha\, \mathrm{d}\boldsymbol{F}_i^\beta\, \mathrm{d}z_i^\beta\,, \tag{8.37}$$

where the normalising constant γ_{ijk} is given by

$$\gamma_{ijk} = \mathcal{N}\left(\mu_\mathrm{f}^k + \mu_\mathrm{f}^j \mid \hat{\boldsymbol{F}}_i, \left[\Sigma^{\hat{F}_i} + \Sigma_\mathrm{ff}^k + \Sigma_\mathrm{ff}^j\right]\right)\,. \tag{8.38}$$

and the combined parameters $\boldsymbol{\eta}^{ijk}$ and $\underline{\boldsymbol{\Lambda}}^{ijk}$ are defined as they are in equations 7.56 and 7.57. We label the integral in equation 8.37 by I_{ijk}, i.e.,

$$I_{ijk} = \iiiint [1 + \xi(z_i^\alpha, z_i^\beta)]\pi(z_i^\alpha, z_i^\beta)\psi(z_i^\alpha, \boldsymbol{F}_i^\alpha)\psi(z_i^\beta, \boldsymbol{F}_i^\beta) \times \\ \tilde{\mathcal{N}}(\boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{\eta}^{ijk}, \underline{\boldsymbol{\Lambda}}^{ijk})\, \mathrm{d}\boldsymbol{F}_i^\alpha\, \mathrm{d}z_i^\alpha\, \mathrm{d}\boldsymbol{F}_i^\beta\, \mathrm{d}z_i^\beta\,. \tag{8.39}$$

As in the previous case of deriving the single-constituent likelihood, the integral I_{ijk} cannot be evaluated analytically. Instead, we again use Monte Carlo integration to calculate this. The Gaussian density $\mathcal{N}(\boldsymbol{F}_i^\alpha, z_i^\alpha, \boldsymbol{F}_i^\beta, z_i^\beta \mid \boldsymbol{\eta}^{ijk}, \underline{\boldsymbol{\Lambda}}^{ijk})$ is normalised to integrate to unity. Thus, given a set of N samples $\{\boldsymbol{F}_{i,j}^\alpha, z_{i,j}^\alpha, \boldsymbol{F}_{i,j}^\beta, z_{i,j}^\beta \mid j = 1 \ldots N\}$

sampled from this density, the integral I_{ijk} is given by

$$I_{ijk} = \frac{\sum_{j=1}^{N}[1 + \xi(z_{i,j}^{\alpha}, z_{i,j}^{\beta})]\pi(z_{i,j}^{\alpha}, z_{i,j}^{\beta})\psi(z_{i,j}^{\alpha}, \boldsymbol{F}_{i,j}^{\alpha})\psi(z_{i,j}^{\beta}, \boldsymbol{F}_{i,j}^{\beta})}{N}\,, \tag{8.40}$$

i.e., the fraction of samples that obey both the sorting condition and the boundary priors, weighted by the redshift correlation function. The two-constituent likelihood is then given by

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^{\alpha}, \boldsymbol{w}^{\beta}, \mathcal{C}_i = 2) &= \mathcal{A}_2 \sum_j \sum_k w_j^{\alpha} w_k^{\beta} I_{ijk} \gamma_{ijk} \\ &\equiv \mathcal{L}_i^2(\boldsymbol{w}^{\alpha}, \boldsymbol{w}^{\beta})\,, \end{aligned}$$

where we use the shortened notation $\mathcal{L}_i^2(\boldsymbol{w}^{\alpha}, \boldsymbol{w}^{\beta})$ to refer to the two-constituent likelihood for source i.

8.1.6 Blending probabilities

Finally, to evaluate the posterior, we need the probabilities for the number of components. In principle, these can be specified *a priori*, e.g., by deriving blending probabilities from morphological information. Here, however, we compute the approximation that these probabilities are fixed using the training set weights $\boldsymbol{w}^{\mathrm{tr}}$ that are otherwise discarded. The single-component probability can then be calculated using Bayes rule by explicitly including the conditioning on $\boldsymbol{w}^{\mathrm{tr}}$ that was suppressed previously, giving

$$P(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\mathrm{tr}}) \propto P(\mathcal{C}_i = 1) P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^{\mathrm{tr}}, \mathcal{C}_i = 1) \tag{8.41}$$

where we have assumed that the prior on the number of components is independent of the test set weights. The evidence term $P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^{\mathrm{tr}}, \mathcal{C}_i = 1)$ is then equal to the single-constituent likelihood defined in equation 8.23, though now conditioned on the training set weights $\boldsymbol{w}^{\mathrm{tr}}$ rather than the parameters $\boldsymbol{w}^1$. Thus, using the result from equation 8.31, the one-constituent probability is given by

$$\begin{aligned} P(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\mathrm{tr}}) &\propto \mathcal{A}_1 P(\mathcal{C}_i = 1) \sum_k w_k^{\mathrm{tr}} \gamma_{ik} I_{ik} \\ &\equiv \tilde{P}(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\mathrm{tr}})\,. \end{aligned} \tag{8.42}$$

Similarly, the two-component probability is given by evaluating the two-constituent likelihood at the training set weights, giving

$$\begin{aligned} P(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr}) &\propto P(\mathcal{C}_i = 2) P(\hat{\boldsymbol{F}}_i \mid \mathcal{C}_i = 2, \boldsymbol{w}^{\rm tr}) \\ &\propto \mathcal{A}_2 P(\mathcal{C}_i = 2) \sum_j \sum_k w_j^{\rm tr} w_k^{\rm tr} I_{ijk} \gamma_{ijk} \\ &\equiv \tilde{P}(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr}) \end{aligned} \tag{8.43}$$

The constant of proportionality for these terms cannot be ignored as it varies with sample. However, we can calculate this by assuming that $\mathcal{C}_i = 1$ and $\mathcal{C}_i = 2$ are the only possible models. The constant of proportionality is then given by

$$\mathcal{Z}_i = \tilde{P}(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr}) + \tilde{P}(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr})$$

The probabilities for one and two components are then given by

$$\begin{aligned} P(\mathcal{C}_i = 1 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr}) &= \frac{\mathcal{A}_1 P(\mathcal{C}_i = 1)}{\mathcal{Z}_i} \sum_k w_k^{\rm tr} \gamma_{ik} I_{ik} \\ &\equiv \mathcal{P}_i^1 \end{aligned}$$

and

$$\begin{aligned} P(\mathcal{C}_i = 2 \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^{\rm tr}) &= \frac{\mathcal{A}_2 P(\mathcal{C}_i = 2)}{\mathcal{Z}_i} \sum_j \sum_k w_j^{\rm tr} w_k^{\rm tr} I_{ijk} \gamma_{ijk} \\ &\equiv \mathcal{P}_i^2 \end{aligned}$$

respectively, where we have defined the shortened notation $\mathcal{P}_i^1$ and $\mathcal{P}_i^2$.

8.1.7 Final posterior

By inserting the above quantities, we find the final expression for the posterior to be

$$\begin{aligned} P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}) \propto \prod_i^N \Big[&\mathrm{Dir}(\boldsymbol{w}^1 \mid \boldsymbol{a}^1) \mathcal{L}_i^1(\boldsymbol{w}^1) \mathcal{P}_i^1 + \\ &\mathrm{Dir}(\boldsymbol{w}^\alpha \mid \boldsymbol{a}^\alpha) \mathrm{Dir}(\boldsymbol{w}^\beta \mid \boldsymbol{a}^\beta) \mathcal{L}_i^2(\boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \mathcal{P}_i^2 \Big] \end{aligned} \tag{8.44}$$

8.2 Histogram model

The only way for the above model to fit each of the redshift distributions is to reweight the existing Gaussian components. The flexibility of the model therefore depends on

the marginal redshift distributions of these components. For example, if these are too wide, the model will be unable to fit narrower features in the test-set redshift distributions.

In order to allow the model more flexibility in the redshift distribution fitting, we can constrain the marginal redshift distributions of each component before fitting the mixture to the test set. We specify these marginal redshift distributions to be uniform bins within a fixed range, so that the redshift distribution inferred by the mixture is a flexible piecewise constant model, i.e., a histogram. We start developing the prior by marginalising over components to give

$$P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}) = \psi(\boldsymbol{F}_i) \sum_k P(\boldsymbol{F}_i, z_i, k \mid \boldsymbol{w}) \,. \tag{8.45}$$

where $\psi(\boldsymbol{F}_i)$ is the boundary prior. This boundary prior now only applies to the fluxes, as we assume that the redshift histogram bin edges are defined such that they enforce positivity. Separating using product rule, this becomes

$$\begin{aligned} P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}) &= \psi(\boldsymbol{F}_i) \sum_k P(k \mid \boldsymbol{w}) P(\boldsymbol{F}_i, z_i \mid k) \\ &\equiv \psi(\boldsymbol{F}_i) \sum_k w_k P(\boldsymbol{F}_i, z_i \mid k) \,, \end{aligned} \tag{8.46}$$

where we have inserted the definition of the weights as the probability of each component.

We now come to specifying the prior while constraining the marginal flux and redshift distributions of each component to be a multivariate Gaussian and histogram bin respectively. In order to allow the likelihood terms to be derived analytically, we make the assumption of conditional independence; given a component k, we assume that the flux and redshift are independent. Separating the right-hand side of equation 8.46 using product rule and utilising this assumption, the prior can be written as

$$P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}) = \psi(\boldsymbol{F}_i) \sum_k w_k P(\boldsymbol{F}_i \mid k) P(z_i \mid k) \,. \tag{8.47}$$

Finally, we insert the desired marginal distributions, giving

$$P(\boldsymbol{F}_i, z_i \mid \boldsymbol{w}) = \psi(\boldsymbol{F}_i) \sum_k w_k \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) \frac{\Theta(z_i - z_k^{\mathrm{lo}}) \Theta(z_k^{\mathrm{hi}} - z_i)}{z_k^{\mathrm{hi}} - z_k^{\mathrm{lo}}} \,. \tag{8.48}$$

where z_k^{lo} and z_k^{hi} are the lower and upper boundaries of the redshift bin corresponding to component k respectively, and $\Theta(\dots)$ is the Heaviside step function.

If each redshift bin is associated with a single component, the height of that bin is $w_k/(z_k^{\rm hi} - z_k^{\rm lo})$. In general, however, more complicated flux-redshift relations can be captured by associating each redshift bin with several components. This is equivalent to giving each redshift bin its own Gaussian mixture model with a small number of components. In this case, the height of the redshift bin j is given by $\sum_{k\in\{k\}_j} w_k/(z_k^{\rm hi} - z_k^{\rm lo})$, where $\{k\}_j$ is the set of components associated with redshift bin j.

8.2.1 Training the Gaussian-constant mixture model

As a simple modification to standard GMMs, the Gaussian-constant mixture model can be easily trained. First, the training set redshifts are binned into a histogram of N_z bins. Each redshift bin b has a height h_b, where the heights have been renormalised so that $\sum_b h_b = 1$.

The fluxes of the training set galaxies within each redshift bin are then fit with a K_b-component GMM. The result is a set of weights $\{\omega_\kappa \mid \kappa = 1 \dots K_b\}$, means $\{\boldsymbol{\mu}_\kappa \mid \kappa = 1 \dots K_b\}$ and covariances $\{\boldsymbol{\Sigma}_\kappa \mid \kappa = 1 \dots K_b\}$ for each redshift bin. The sets of means and covariances can simply be combined into sets of $N_z \times K_b$ component parameters for the full Gaussian-constant mixture model. However, the weights must first be normalised by the bin heights, so that

$$w_k = h_k \omega_k \,, \tag{8.49}$$

where h_k is the height of the redshift bin corresponding to component k, and $\omega_k \equiv \omega_{\kappa,b}$ is the GMM weight of the component indexed by k in the full model and by κ in the GMM fitted to redshift bin b. Since these were both normalised to unity, the final model weights are correctly normalised too[1].

8.2.2 Single-constituent likelihood

In order to insert this Gaussian-constant mixture prior into the above model in place of the GMM, we need to recalculate the $(I\gamma)_{ik} \equiv I_{ik}\gamma_{ik}$ and $(I\gamma)_{ijk} \equiv I_{ijk}\gamma_{ijk}$ terms; that is, the product of the likelihood and the prior, marginalised over the latent flux and redshift. The prior normalisations $\mathcal{A}_1$ and $\mathcal{A}_2$ also must be recalculated using samples from the new prior instead of the previous full-GMM prior. Since the structure of the

[1] $\sum_k w_k = \sum_k h_k \omega_k = \sum_b h_b \sum_\kappa \omega_{\kappa,b} = \sum_b h_b = 1$

two models is the same except for the new prior, the other results all hold once these are replaced. We first calculate the one-component term, defined as

$$(I\gamma)_{ik} \equiv \iint P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i, \mathcal{C}_i = 1) P(\boldsymbol{F}_i, z_i \mid k) \, \mathrm{d}\boldsymbol{F}_i \, \mathrm{d}z_i \,. \tag{8.50}$$

Inserting the definition of each term, this becomes

$$(I\gamma)_{ik} = \iint \psi(\boldsymbol{F}_i) \mathcal{N}(\boldsymbol{F}_i \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) \frac{\Theta(z_i - z_k^{\mathrm{lo}}) \Theta(z_k^{\mathrm{hi}} - z_i)}{z_k^{\mathrm{hi}} - z_k^{\mathrm{lo}}} \, \mathrm{d}\boldsymbol{F}_i \, \mathrm{d}z_i \,. \tag{8.51}$$

As we did before, the two Gaussian densities can be combined into a single multivariate Gaussian. This is a standard result, as the dimensionalities of both densities match, giving

$$\mathcal{N}(\boldsymbol{F}_i \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) = \gamma_{ik}^* \, \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}^{ik}, \underline{\boldsymbol{\Sigma}}^{ik}) \,, \tag{8.52}$$

where the constant of proportionality is given by

$$\gamma_{ik}^* = \mathcal{N}(\boldsymbol{\mu}_k \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}_k + \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \,, \tag{8.53}$$

and the new parameters $\boldsymbol{\mu}^{ik}$ and $\underline{\boldsymbol{\Sigma}}^{ik}$ are given by the results in section 8.1.8 in Petersen and Pedersen (2014). The redshift histogram density is normalised to integrate to unity, i.e.,

$$\int \frac{\Theta(z_i - z_k^{\mathrm{lo}}) \Theta(z_k^{\mathrm{hi}} - z_i)}{z_k^{\mathrm{hi}} - z_k^{\mathrm{lo}}} \, \mathrm{d}z_i = 1 \,, \tag{8.54}$$

provided that every value contained within the bin is positive. The edges of the redshift histogram bin can be chosen freely, so we assume this to be the case. Thus, $(I\gamma)_{ik}$ becomes

$$(I\gamma)_{ik} = \mathcal{N}(\boldsymbol{\mu}_k \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}_k + \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \int \psi(\boldsymbol{F}_i) \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}^{ik}, \underline{\boldsymbol{\Sigma}}^{ik}) \, \mathrm{d}\boldsymbol{F}_i \tag{8.55}$$

As before, the integral here can be calculated numerically through Monte Carlo integration to be

$$\int \psi(\boldsymbol{F}_i) \mathcal{N}(\boldsymbol{F}_i \mid \boldsymbol{\mu}^{ik}, \underline{\boldsymbol{\Sigma}}^{ik}) \, \mathrm{d}\boldsymbol{F}_i \approx \frac{\sum_{j=1}^{N} \psi(\boldsymbol{F}_j)}{N} \,, \tag{8.56}$$

given a set of N samples $\{\boldsymbol{F}_j \mid j = 1 \ldots N\}$ drawn from the density in the integrand.

8.2.3 Two-constituent likelihood

We now calculate the two-component term $(I\gamma)_{ijk}$. This is defined to be

$$(I\gamma)_{ijk} \equiv \iiiint P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta) P(\boldsymbol{F}_i^\alpha, z_i^\alpha \mid j) P(\boldsymbol{F}_i^\beta, z_i^\beta \mid k) \pi(z_i^\alpha, z_i^\beta) \, \mathrm{d}\boldsymbol{F}_i^\alpha \, \mathrm{d}z_i^\alpha \, \mathrm{d}\boldsymbol{F}_i^\beta \, \mathrm{d}z_i^\beta \,, \tag{8.57}$$

where we have neglected the correlation function $\xi(z_1, z_2)$ to allow analytic redshift integrations. Inserting the definitions of each of the terms in the integrand, this becomes

$$\begin{aligned}(I\gamma)_{ijk} = \iint & \mathcal{N}(\boldsymbol{F}_i^\alpha + \boldsymbol{F}_i^\beta \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \mathcal{N}(\boldsymbol{F}_i^\alpha \mid \boldsymbol{\mu}_j, \underline{\boldsymbol{\Sigma}}_j) \mathcal{N}(\boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) \psi(\boldsymbol{F}_i^\alpha) \psi(\boldsymbol{F}_i^\beta) \times \\ & \left[\iint \frac{\Theta(z_i^\alpha - z_j^{\mathrm{lo}}) \Theta(z_j^{\mathrm{hi}} - z_i^\alpha)}{z_j^{\mathrm{hi}} - z_j^{\mathrm{lo}}} \, \frac{\Theta(z_i^\beta - z_k^{\mathrm{lo}}) \Theta(z_k^{\mathrm{hi}} - z_i^\beta)}{z_k^{\mathrm{hi}} - z_k^{\mathrm{lo}}} \times \right. \\ & \left. \pi(z_i^\alpha, z_i^\beta) \, \mathrm{d}z_i^\alpha \, \mathrm{d}z_i^\beta \right] \mathrm{d}\boldsymbol{F}_i^\alpha \, \mathrm{d}\boldsymbol{F}_i^\beta \,.\end{aligned} \tag{8.58}$$

Because of the sorting condition, the result of the redshift integrals depends on the values of j and k. Assume that the bins are ordered so that increasing index corresponds to increasing redshift. Then, the redshift integral is given by

$$Y_{jk} \equiv \iint \frac{\Theta(z_i^\alpha - z_j^{\mathrm{lo}}) \Theta(z_j^{\mathrm{hi}} - z_i^\alpha)}{z_j^{\mathrm{hi}} - z_j^{\mathrm{lo}}} \, \frac{\Theta(z_i^\beta - z_k^{\mathrm{lo}}) \Theta(z_k^{\mathrm{hi}} - z_i^\beta)}{z_k^{\mathrm{hi}} - z_k^{\mathrm{lo}}} \pi(z_i^\alpha, z_i^\beta) \, \mathrm{d}z_i^\alpha \, \mathrm{d}z_i^\beta \,. \tag{8.59}$$

This integral can be evaluated analytically to be

$$Y_{jk} = \begin{cases} 0 & \text{when } j > k \\ \frac{1}{2} & \text{when } j = k \\ 1 & \text{when } j < k \end{cases} \tag{8.60}$$

Inserting this into the equation above, we get

$$\begin{aligned}(I\gamma)_{ijk} = Y_{jk} \iint & \mathcal{N}(\boldsymbol{F}_i^\alpha + \boldsymbol{F}_i^\beta \mid \hat{\boldsymbol{F}}_i, \underline{\boldsymbol{\Sigma}}^{\hat{F}_i}) \mathcal{N}(\boldsymbol{F}_i^\alpha \mid \boldsymbol{\mu}_j, \underline{\boldsymbol{\Sigma}}_j) \mathcal{N}(\boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) \times \\ & \psi(\boldsymbol{F}_i^\alpha) \psi(\boldsymbol{F}_i^\beta) \, \mathrm{d}\boldsymbol{F}_i^\alpha \, \mathrm{d}\boldsymbol{F}_i^\beta\end{aligned} \tag{8.61}$$

We can combine the two prior terms trivially as they are independent, giving

$$\mathcal{N}(\boldsymbol{F}_i^\alpha \mid \boldsymbol{\mu}_j, \underline{\boldsymbol{\Sigma}}_j) \mathcal{N}(\boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_k, \underline{\boldsymbol{\Sigma}}_k) = \mathcal{N}(\boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_{jk}, \underline{\boldsymbol{\Sigma}}_{jk}) \tag{8.62}$$

where

$$\boldsymbol{\mu}_{jk} = \begin{pmatrix} \boldsymbol{\mu}^j \\ \boldsymbol{\mu}^k \end{pmatrix} \qquad \underline{\boldsymbol{\Sigma}}_{jk} = \begin{pmatrix} \underline{\boldsymbol{\Sigma}}_j & 0 \\ 0 & \underline{\boldsymbol{\Sigma}}_k \end{pmatrix} . \tag{8.63}$$

We can then define the corresponding natural parameters as $\underline{\boldsymbol{\Lambda}}_{jk} \equiv \underline{\boldsymbol{\Sigma}}_{jk}^{-1}$ and $\boldsymbol{\eta}_{jk} \equiv \underline{\boldsymbol{\Sigma}}_{jk}^{-1}\boldsymbol{\mu}_{jk}$, the blocks of which we label as

$$\boldsymbol{\eta}_{jk} = \begin{pmatrix} \boldsymbol{\eta}_a^{jk} \\ \boldsymbol{\eta}_b^{jk} \end{pmatrix} \qquad \underline{\boldsymbol{\Lambda}}^{jk} = \begin{pmatrix} \underline{\boldsymbol{\Lambda}}_{aa'}^{jk} & \underline{\boldsymbol{\Lambda}}_{ab'}^{jk} \\ \underline{\boldsymbol{\Lambda}}_{ba'}^{jk} & \underline{\boldsymbol{\Lambda}}_{bb'}^{jk} \end{pmatrix} . \tag{8.64}$$

where all of the blocks of each parameter are of equal size. The remaining two densities can now be combined to give

$$\begin{aligned} (I\gamma)_{ijk} = Y_{jk}\, \mathcal{N}(\boldsymbol{\mu}_j + \boldsymbol{\mu}_k \mid \hat{\boldsymbol{F}}_i, [\underline{\boldsymbol{\Sigma}}^{\hat{F}_i} + \underline{\boldsymbol{\Sigma}}_j + \underline{\boldsymbol{\Sigma}}_k]) \times \\ \iint \mathcal{N}(\boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_{ijk}, \underline{\boldsymbol{\Sigma}}_{ijk}) \psi(\boldsymbol{F}_i^\alpha) \psi(\boldsymbol{F}_i^\beta)\, \mathrm{d}\boldsymbol{F}_i^\alpha\, \mathrm{d}\boldsymbol{F}_i^\beta \end{aligned} \tag{8.65}$$

where

$$\boldsymbol{\eta}_{ijk} = \begin{pmatrix} \boldsymbol{\eta}_a^{jk} + \boldsymbol{\eta}^{F_i} \\ \boldsymbol{\eta}_b^{jk} + \boldsymbol{\eta}^{F_i} \end{pmatrix} \qquad \underline{\boldsymbol{\Lambda}}^{ijk} = \begin{pmatrix} \underline{\boldsymbol{\Lambda}}_{aa'}^{jk} + \underline{\boldsymbol{\Lambda}}^{\hat{F}_i} & \underline{\boldsymbol{\Lambda}}_{ab'}^{jk} + \underline{\boldsymbol{\Lambda}}^{\hat{F}_i} \\ \underline{\boldsymbol{\Lambda}}_{ba'}^{jk} + \underline{\boldsymbol{\Lambda}}^{\hat{F}_i} & \underline{\boldsymbol{\Lambda}}_{bb'}^{jk} + \underline{\boldsymbol{\Lambda}}^{\hat{F}_i} \end{pmatrix} . \tag{8.66}$$

The integral over fluxes can then be approximated through Monte Carlo integration as

$$\iint \mathcal{N}(\boldsymbol{F}_i^\alpha, \boldsymbol{F}_i^\beta \mid \boldsymbol{\mu}_{ijk}, \underline{\boldsymbol{\Sigma}}_{ijk}) \psi(\boldsymbol{F}_i^\alpha) \psi(\boldsymbol{F}_i^\beta)\, \mathrm{d}\boldsymbol{F}_i^\alpha\, \mathrm{d}\boldsymbol{F}_i^\beta \approx \frac{\sum_{j=1}^N \psi(\boldsymbol{F}_{i,j}^\alpha)\psi(\boldsymbol{F}_{i,j}^\beta)}{N} , \tag{8.67}$$

given a set of N samples $\{\boldsymbol{F}_{i,j}^\alpha, \boldsymbol{F}_{i,j}^\beta \mid j = 1 \ldots N\}$ drawn from the Gaussian density in the integrand.

8.3 Testing the models on simulated data

In order to test the two methods above, we used the simulated LSST (Ivezić et al., 2019) and Euclid (Laureijs et al., 2011) datasets described in section 7.3. We simulated 10000 unblended sources and 10000 two-constituent blended sources as inputs to the methods, and assumed a uniform prior on the number of constituents, i.e., $P(\mathcal{C}_i = 1) = P(\mathcal{C}_i = 2) = 0.5$.

We implemented both models using Stan (Carpenter et al., 2017), a probabilistic programming language and inference library. Using Stan, inference can be performed on models using three different methods. Firstly, the maximum *a posteriori* (MAP) parameters can be found using the L-BFGS (Byrd et al., 1995) optimiser. Secondly, the posterior distribution can be sampled using NUTS (Hoffman and Gelman, 2011), an automatically-tuning variety of Hamiltonian Monte Carlo (HMC). Finally, the posterior distribution can be approximated using variational inference (VI), where an uncorrelated multivariate Gaussian approximating distribution is fitted in a transformed parameter space where constrained parameters are transformed to have infinite support (Kucukelbir et al., 2015). See chapter 3 for further discussion of these inference approaches.

The $(I\gamma)_{ik} \equiv I_{ik}\gamma_{ik}$ and $(I\gamma)_{ijk} \equiv I_{ijk}\gamma_{ijk}$ terms and the blending probabilities $\mathcal{P}_i^1$ and $\mathcal{P}_i^2$ are constant w.r.t. the weight vectors, and so can be pre-calculated and cached. We then perform inference on each model using the three inference methods above. The results of these tests on each of the models using the same simulated dataset are detailed below.

Figure 8.3 shows the inferred redshift distributions for the hierarchical GMM method described in section 8.1. The top row shows the inferred distributions for the unblended, low-redshift-blended and high-redshift-blended constituents obtained using HMC. While the overall width of the true distributions has been approximately identified, these results exhibit strong oscillatory behaviour that has not been accounted for. Further work would be required to identify the cause of this effect. The bottom row shows the same results obtained using variational inference. These results are very similar to those obtained by HMC despite variational inference only being an approximate inference method. As a result, variational inference offers a significant computational advantage. After caching several terms as described above at a rate of ≈ 10 sources s^{-1}, VI completed inference in ≈ 1 hr., compared to ≈ 12 hrs. for HMC on this dataset of 20000 sources.

Figure 8.4 shows the inferred redshift distributions for the histogram method described in section 8.2. The restriction in the distributions that can be described by this model mean that less oscillatory behaviour is seen than in the results of the previous model. However, the true distributions are still not successfully recovered to the precision implied by the posterior distributions over bin heights. More work would again be required to understand the cause of this. This figure also shows that the results obtained using variational inference again closely match those obtained using HMC while offering a similar gain in computational efficiency.

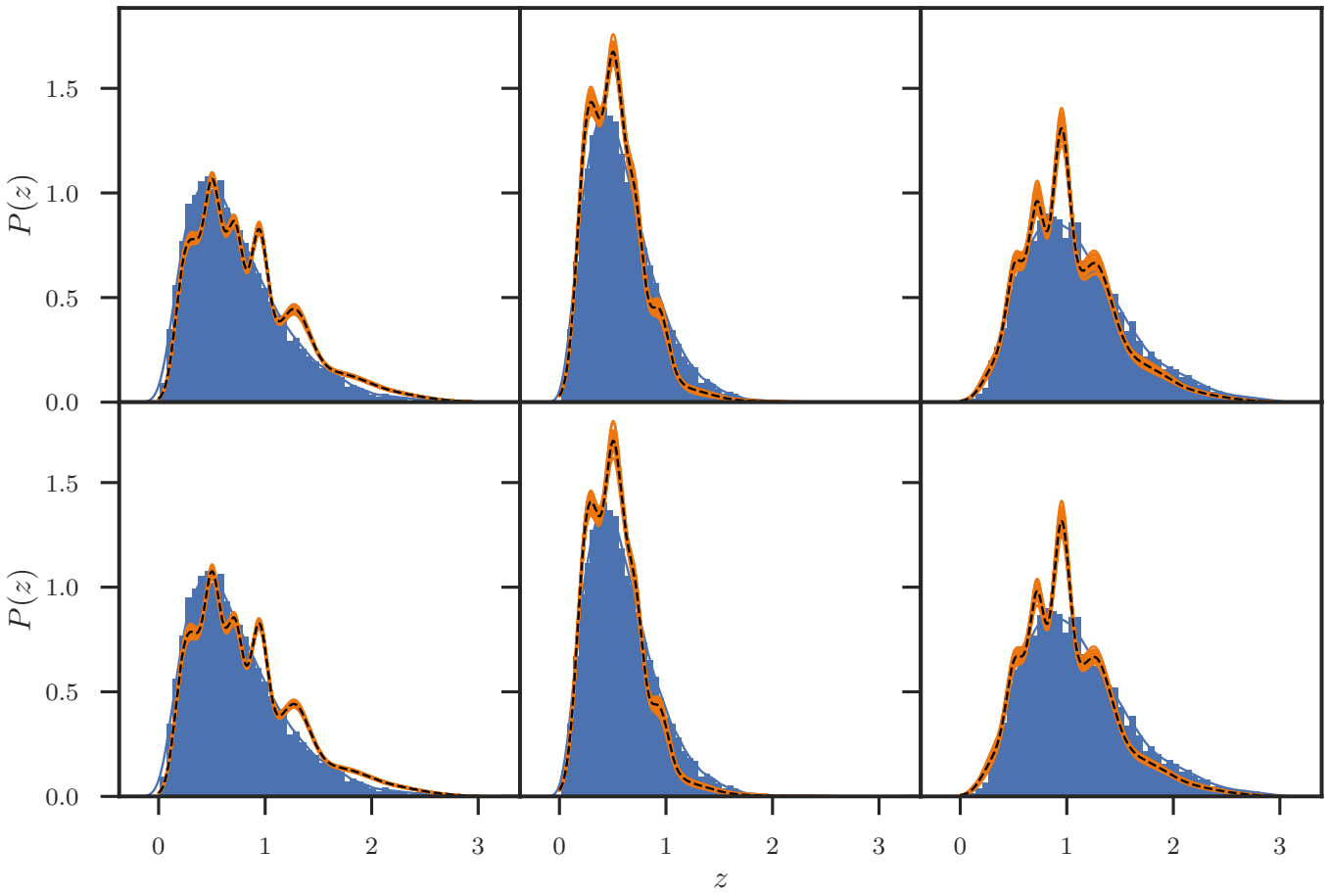

Figure 8.3: Plots showing the inferred redshift distributions from simulated data using the hierarchical GMM method. The blue histograms show the true distributions, while the blue lines shows a kernel density estimate of these distributions to smooth them for comparison with the continuous inferred distributions. The left panels correspond to the unblended sources, the centre panels correspond to the lower-redshift constituent of the blended sources, and the right panels correspond to the higher-redshift constituent of the blended sources. The orange curves in the top row are samples from the exact posterior sampled using HMC, and the orange curves in the bottom row are approximations obtained using variational inference. The black dashed lines in all panels show the maximum *a posteriori* distributions found by the L-BFGS (Byrd et al., 1995) optimiser.

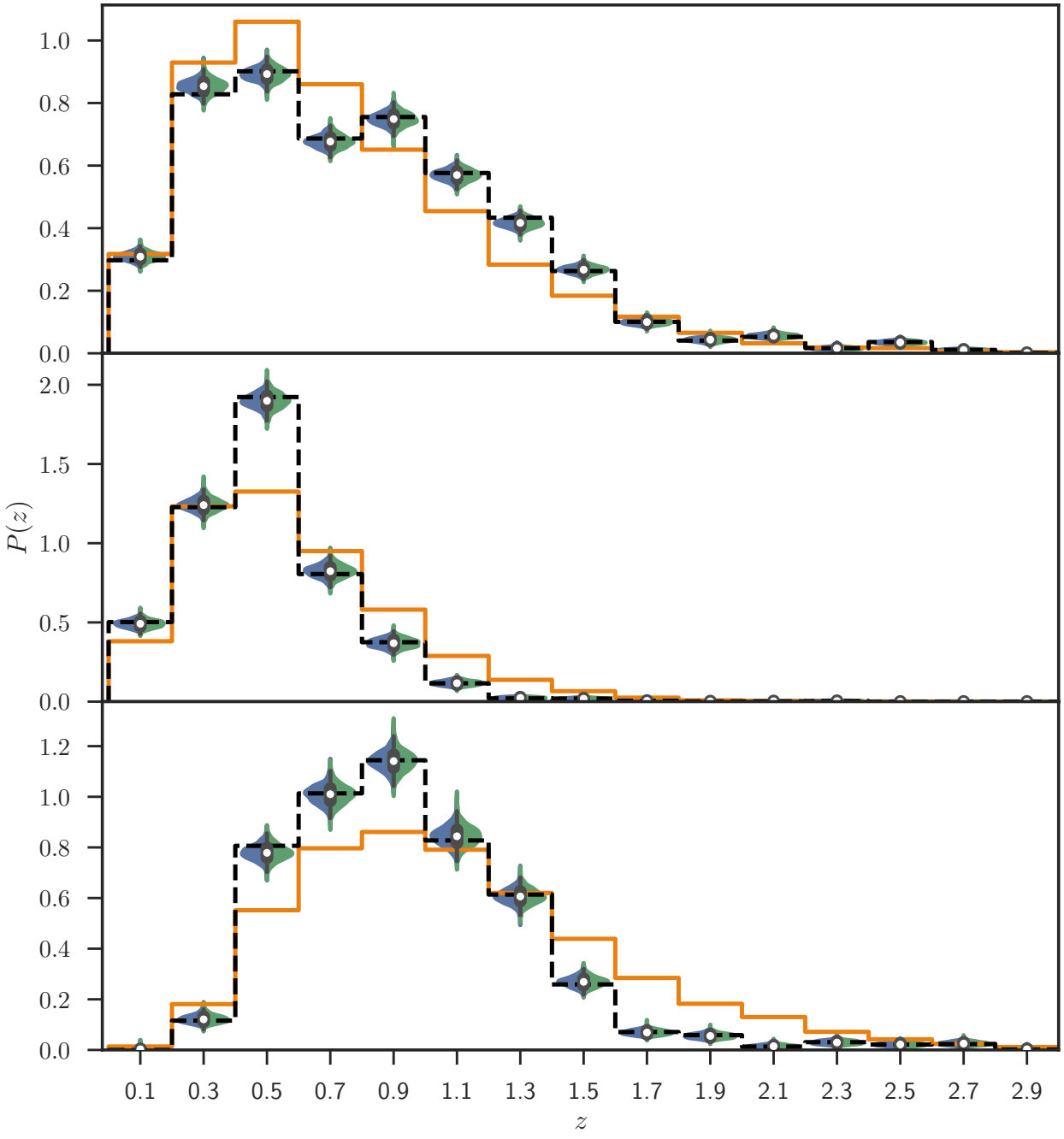

Figure 8.4: Plots showing the inferred redshift distributions from simulated data using the histogram method. The top panel corresponds to the unblended sources, the centre panel corresponds to the lower-redshift constituent of the blended sources, and the bottom panel corresponds to the higher-redshift constituent of the blended sources. The orange histogram shows the true distribution, and the black dashed lines show the maximum *a posteriori* distributions found by the L-BFGS (Byrd et al., 1995) optimiser. The violin plots show the distribution of bin heights inferred from the posterior, with the samples from HMC shown in the blue left-halves and the samples from variational inference shown in the green right-halves.

Both of these models show somewhat promising results that still require some work to improve. It is possible that the approximation of fixed blending probability made in both of these methods is a poor one that negatively affects the results. Further work to relax this assumption through Gibbs sampling is described below.

8.4 Further work

This section describes several possible extensions to improve the above methods.

8.4.1 Gibbs sampling

As described in section 8.1.1, an alternative to the approximation of fixed blending probability made above is Gibbs sampling. Rather than marginalising over the number of constituents $\{\mathcal{C}\}$, we can include these parameters in the joint posterior, i.e., $P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \{\mathcal{C}\} \mid \{\hat{\boldsymbol{F}}\})$. We then write this is terms of two conditional distributions that can be Gibbs sampled. Firstly, we can use Bayes rule and assume i.i.d. data to write the conditional distribution for the weights as

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}, \{\mathcal{C}\}) \propto P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\mathcal{C}\}) \prod_i^N P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i) \tag{8.68}$$

where the sample-i likelihood is given by

$$\begin{aligned} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i) &= \begin{cases} P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^1, \mathcal{C}_i = 1) = \sum_k w_k^1 \gamma_{ik} \\ P(\hat{\boldsymbol{F}}_i \mid \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \mathcal{C}_i = 2) = \sum_j \sum_k w_j^\alpha w_k^\beta I_{ijk} \gamma_{ijk} \end{cases} \\ &\equiv \mathcal{L}_i^{\mathcal{C}_i}(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \end{aligned} \tag{8.69}$$

as shown in section 8.1. We assume that the weight priors are independent of each other and of the number of components for each sample, i.e.,

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\mathcal{C}\}) \equiv P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \equiv P(\boldsymbol{w}^1) P(\boldsymbol{w}^\alpha) P(\boldsymbol{w}^\beta) \tag{8.70}$$

We also assume that each of these are Dirichlet distributed as above, i.e.,

$$\begin{aligned} P(\boldsymbol{w}^1) &= \mathrm{Dir}(\boldsymbol{w}^1 \mid \boldsymbol{a}^1) \\ P(\boldsymbol{w}^\alpha) &= \mathrm{Dir}(\boldsymbol{w}^\alpha \mid \boldsymbol{a}^\alpha) \\ P(\boldsymbol{w}^\beta) &= \mathrm{Dir}(\boldsymbol{w}^\beta \mid \boldsymbol{a}^\beta) \end{aligned} \tag{8.71}$$

Thus, the conditional distribution for the weights is given by

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}, \{\mathcal{C}\}) \propto \mathrm{Dir}(\boldsymbol{w}^1 \mid \boldsymbol{a}^1)\mathrm{Dir}(\boldsymbol{w}^\alpha \mid \boldsymbol{a}^\alpha)\mathrm{Dir}(\boldsymbol{w}^\beta \mid \boldsymbol{a}^\beta) \prod_i^N \mathcal{L}_i^{\mathcal{C}_i}(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \tag{8.72}$$

Next, we derive the conditional distribution for the number of components. Again assuming independence, this can be written

$$P(\{\mathcal{C}\} \mid \{\hat{\boldsymbol{F}}\}, \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) = \prod_i P(\mathcal{C}_i \mid \hat{\boldsymbol{F}}_i, \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \tag{8.73}$$

Applying Bayes rule, this becomes

$$P(\{\mathcal{C}\} \mid \{\hat{\boldsymbol{F}}\}, \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \propto \prod_i P(\hat{\boldsymbol{F}}_i \mid \mathcal{C}_i, \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) P(\mathcal{C}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \tag{8.74}$$

Finally, we assume that the prior probability of blending is independent of the weights, e.g., it's set by the proportion of sources we expect to be blended in the sample, or whether sample i is in a cluster or the field. Thus, $P(\mathcal{C}_i \mid \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) = P(\mathcal{C}_i)$. Equation 8.74 can then be written using the likelihood result from above as

$$P(\{\mathcal{C}\} \mid \{\hat{\boldsymbol{F}}\}, \boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \propto \prod_i P(\mathcal{C}_i) \mathcal{L}_i^{\mathcal{C}_i}(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta) \tag{8.75}$$

8.4.2 Relaxing the assumption of conditional independence in the histogram model with copulas

In the derivation of the histogram model in section 8.2, we made use of the assumption of conditional independence where, given a component k, the flux and redshift were assumed independent. This allowed us to analytically specify a prior distribution where the marginal redshift distribution was constrained to be a histogram bin. In principle, however, it is possible to construct probability distributions by separately specifying the desired marginal distributions and correlation structure by using *copulas* (e.g., Trivedi et al., 2007).

Copulas make use of a property known as the *probability integral transform*. Let $\mathrm{CDF}(x)$ be the cumulative distribution function corresponding to the a distribution with PDF $P(x)$. Then, given samples drawn from this distribution $x \sim P(x)$, the CDF can transform these samples by $y = \mathrm{CDF}(x)$ so that the transformed parameter y is uniformly distributed, i.e., $y \sim U(0, 1)$. The inverse CDF can be used to make the

inverse transformation, i.e., $\mathrm{CDF}^{-1}(y) = x$.

This probability integral transform can be used to construct a copula as follows. First, we sample $\boldsymbol{a} \sim P(\boldsymbol{x})$ from a probability distribution with the desired correlations. Next, this is transformed by $\boldsymbol{b} = \mathrm{CDF}(\boldsymbol{a})$ to obtain vectors with elements that are correlated but have uniform marginal distributions. We then make a final transformation $c_i = \mathrm{CDF}_i^{-1}(b_i)$ where $\mathrm{CDF}_i^{-1}(\dots)$ is the inverse CDF of the desired marginal distribution of dimension i. The vector $\boldsymbol{c}$ will then be distributed with a correlated joint distribution but with the specified marginal distributions as desired.

While this is possible in principle, further work would be required to determine whether inference could still be performed efficiently for a copula model.

8.4.3 Fully hierarchical model over components

A major simplifying assumption made above is that the mixture parameters $\{\mu\}$ and $\{\Sigma\}$ are initially fitted to a training set as described in section 8.1 and then held constant. This approach has some significant computational benefits. By assuming these parameters to be fixed and varying only the weights, the resulting posterior distribution is easy to explore. If these parameters are instead allowed to vary, inference can become extremely difficult. Mixture components can be exchanged, meaning that the resulting posterior is highly multimodal. While enforcing an ordering in the components would remove this degeneracy, e.g., by including a prior that $\boldsymbol{w}_i \leq \boldsymbol{w}_{i+1}$, Bayesian mixture models with strongly-overlapping components still exhibit pathological posterior geometries that can render sampling extremely inefficient (Betancourt, 2017).

Nevertheless, if further work were able to overcome these computational difficulties, a fully hierarchical approach which jointly inferred the mixture parameters could have some advantages. By giving the model freedom to vary the mixture parameters, a set of Gaussian components that fits both the training and test sets could be found, rather than assuming that a set of components fitted only on the training set would be applicable to the test set. To do this, we would need to condition on both the training and test set data in the posterior. The posterior over the quantities we're interested in could then be obtained by marginalising over the full hierarchical posterior like

$$P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta \mid \{\hat{\boldsymbol{F}}\}, \{\hat{\boldsymbol{F}}^{\mathrm{tr}}\}, \{\hat{z}^{\mathrm{tr}}\}) = \iiint P(\boldsymbol{w}^1, \boldsymbol{w}^\alpha, \boldsymbol{w}^\beta, \boldsymbol{w}^{\mathrm{tr}}, \{\boldsymbol{\mu}_k\}, \{\underline{\boldsymbol{\Sigma}}_k\} \mid \{\hat{\boldsymbol{F}}\}, \{\hat{\boldsymbol{F}}^{\mathrm{tr}}\}, \{\hat{z}^{\mathrm{tr}}\}) \, \mathrm{d}\boldsymbol{w}^{\mathrm{tr}} \, \mathrm{d}\{\boldsymbol{\mu}_k\} \, \mathrm{d}\{\underline{\boldsymbol{\Sigma}}_k\} \,. \tag{8.76}$$

This posterior could then be developed to allow Gibbs sampling as described in section 8.4.1. Alternatively, the training set could be fitted separately with a GMM as above, and the resulting weights, means and covariances used to calculate the fixed blending probability only.

8.5 Conclusions

In order to use photometric galaxy surveys to do cosmology, it is necessary to estimate the redshift distribution of the observed sources. Photometric redshift methods designed for inferring these population distributions are therefore an important part of the cosmological analysis pipeline. If uncertainties in cosmological parameters are to be an accurate reflection of our state of knowledge, uncertainties from all parts of the analysis should be correctly propagated, meaning that these redshift distributions should have associated errors.

This chapter extends the GMM method of chapter 7 into a Bayesian hierarchical model that can be used to infer posterior distributions over redshift distributions for a population of possibly blended sources. This is done by modelling unblended sources, the lower-redshift constituent of blended sources, and the upper-redshift constituent of blended sources as three separate mixture models. These mixtures share means and covariances that are initially fitted to an unblended training set, but vary in their weights. This chapter also describes an alternative model where each mixture component is a multivariate Gaussian in fluxes and a small bin in redshift. The result of this is that the overall mixture describes that redshift distribution as a histogram.

We test these models using a dataset of simulated LSST and Euclid sources, inferring posterior distributions using both Hamiltonian Monte Carlo and variational inference. While the GMM model recovers the overall width of each distribution, the resulting distributions show an unexplained oscillatory behaviour. The histogram model prevents this oscillatory behaviour but does not recover the distributions well. These results suggest that more work would be required before the methods could be applied to the analysis of cosmological galaxy surveys. However, our results indicate that variational inference offers a good approximation to Hamiltonian Monte Carlo for this problem while providing a substantial increase in computational efficiency, an important trait for application to large future datasets.

The models in this chapter are based on the Leistedt et al. (2016) hierarchical Bayesian model for inferring redshift distributions. The biggest difference between these approaches is that the models in this chapter show a generalisation to blended

sources. However, there are also additional differences beyond this. Firstly, the models in this chapter are machine learning-based, learning the flux-redshift distribution from a labelled training set. In contrast, the model in Leistedt et al. (2016) is template-based. In addition, inference in the Leistedt et al. (2016) model is performed using a custom Gibbs sampling approach. Instead, due to the choice of model and approximations made throughout, the models in this chapter are amenable to being implemented in general purpose inference software such as Stan (Carpenter et al., 2017), allowing both Hamiltonian Monte Carlo and variational inference.

Clustering redshifts (e.g., Newman, 2008; Schmidt et al., 2013; Ménard et al., 2013) are another method for inferring redshift distributions of photometric samples. These methods utilise only the positional information from photometric observations rather than the colour information. The position of sources is likely to have a small error provided that a source has been successfully observed, while faint sources may have large photometric errors that make colour-based photometric redshifts inaccurate. In addition, clustering-based methods require a population of sources with known redshifts that overlap with the target population in both redshift and position on the sky. In some sense, this is a stronger requirement than a training set that covers the full range of *redshifts* only. However, the overlap population need not be representative of the target sources, a requirement of colour-based machine learning photometric redshift methods such as the models presented in this chapter.

Clustering information and colour information are complementary and thus can be combined. The model described in Sánchez and Bernstein (2019) and Alarcon et al. (2019) is an extension of the Leistedt et al. (2016) model to also include clustering data. Such an extension could also be performed for the models presented here, though the clustering would also need to be generalised to the case of blended sources.

The models in this chapter are also similar in spirit to the DIR calibration method of Lima et al. (2008). In that approach, the photometric redshift distribution is given by the redshift distribution of a training set of spectroscopic sources where each source is weighted. This weight is give such that the flux distribution of the weighted sample matches that of the photometric test set, as measured using a nearest neighbour method. The models presented in this chapter also involve inferring a set of weights. However, since these weights correspond to mixture components rather than each training galaxy, their number is significantly smaller than in the DIR method. This enables inference methods such as MCMC, in contrast to the bootstrapping method utilised for uncertainties in the DIR approach (e.g., Joudaki et al., 2019). This modelling approach also enables the extension to blended sources that motivated this work.

Part III

Conclusions

Chapter 9

Thesis Summary and Future Work

Despite the substantial successes of the ΛCDM cosmological model, several open problems and tensions remain. Many of these open problems could be addressed by future cosmological galaxy surveys, dedicated experiments which systematically image large volumes of the universe. In order to place constraints on cosmological parameters and conduct tests of ΛCDM which could lead to progress on these outstanding problems, galaxy surveys require knowledge of the redshifts of the sources that they observe.

Due to limitations in the telescope time and the large depths these surveys will observe to, spectroscopic measurements of redshifts are infeasible in practice. As a result, photometric redshifts will continue to be a vital tool in the cosmological analysis of galaxy surveys. Photometric redshifts are statistical methods that estimate the redshifts of sources from photometry, a handful of flux measurements obtained from images using a small number of broadband filters. In addition some photometric methods are designed for inferring the redshift distribution of a population of sources from their photometry.

Future galaxy surveys will observe to greater depths than previous surveys have, driving an increase in the precision of the resulting cosmological constraints. However, this increase in depth also increases the fraction of sources that will overlap along the line of sight, an effect known as blending. For future surveys such as LSST, around half of sources will be blended to some degree. Methods to separate blended sources, known as deblending, have therefore been developed. However, completely propagating all uncertainties from measurements made on deblended sources, including all correlations, can be difficult.

The work in this thesis take a different approach to deblending. We develop photometric redshift methods that infer the redshifts from blended data directly. By framing the problem in this manner, all uncertainties can be accounted for simply.

In chapter 6, we generalise existing Bayesian template-based photometric redshift methods to the case of blended sources. We derive an expression for the posterior distribution over the redshift and magnitude for each constituents galaxy in a blended source with an arbitrary number of constituents. Using Bayesian model comparison, we are also able to determine probabilities for the number of constituents in each source. We test this method on both simulated data and real spectroscopically confirmed blended data with known redshifts. We also test the method on simulated partially-blended sources where only some bands are blended, finding that this reduces the fraction of outliers.

In chapter 7, we tackle the same physical problem using the other main type of photometric redshifts, empirical methods. We model the joint flux-redshift distribution of galaxies as a Gaussian mixture model (GMM), and fit this model to a training set of unblended sources. Given this model, we derive both the redshift posterior distributions and the Bayesian evidences for of one- and two-constituent sources. Due to our choice of using GMMs, the posterior distributions can be efficiently sampled without MCMC techniques. In addition, Bayesian model comparison can be used to provide probabilities for the number of sources much faster than the template-based method in the previous chapter, an important property for the large datasets of future surveys. As before, we test this method on both simulated and real data. We find that redshifts are well recovered in regions where sources are well represented in the training set, though the results are expectedly worse for sources fainter than the depth of the training set.

Finally, chapter 8 extends the GMM of the previous chapter to a hierarchical model in order to infer posterior distributions over the redshift distributions of a population of possibly blended sources. To do this, we model the distributions of the unblended galaxies, the lower-redshift constituent of blended sources, and the higher-redshift constituent of blended sources as three separate GMMs that share the sets of means and covariances but differ in their weights. We test this method on simulated sources using both Hamiltonian Monte Carlo and variational inference techniques, finding that the overall width of the distributions is well recovered, though the distributions in detail display oscillatory behaviour that would require more work to understand. In addition, we also test a mixture model where each component is a multivariate Gaussian in the fluxes but is a uniform bin in redshifts. The resulting output is therefore a posterior distribution over redshift distributions histograms. This stops the oscillatory behaviour of the previous model, though the distributions are not well recovered to the precision suggested by the posterior distributions. Chapter 8 also describes several possible extensions to the models described that could improve their results.

Table 9.1: A summary of the research results in this thesis.

	Chapter 6	Chapter 7	Chapter 8
Method type	Template-based	Empirical (GMM)	Empirical (GMM)
Inference target	Per-object redshift PDFs and Bayesian evidence for each number of constituents	Per-object redshift PDFs and Bayesian evidence for each number of constituents	Population redshift distribution PDFs
Inference method	Nested sampling (MultiNest)	Rejection sampling	HMC and VI
Redshift scatter on simulated blended data	$\sigma_{\mathrm{RMS}} = 0.163$	$\sigma_{\mathrm{RMS}} = 0.105$	N/A
Blends identified in simulated blended data	92.7%	92.4%	N/A
Redshift scatter on GAMA blended sources catalogue	$\sigma_{\mathrm{RMS}} = 0.156$	$\sigma_{\mathrm{RMS}} = 0.091$	N/A
Blends identified in GAMA blended sources catalogue	71.6%	33.4%	N/A
Computation time	$\sim$ 2 mins per source	$\sim$ 0.1 secs per source	$\sim$ 0.1 secs caching per source + 12 (HMC) / 1 (VI) hrs sampling for 20000 sources

We summarise and compare the results of each research chapter in Table 9.1. The methods in chapters 6 and 7 tackle the same per-source photometric redshift problem generalised to blended sources using different methods. These methods are then tested on both simulated and real datasets. It is worth noting that both methods achieved a slightly lower scatter on the real blended sources from the GAMA blended sources catalogue (Holwerda et al., 2015) than the simulated LSST-like data. A likely reason for this is that the GAMA dataset represents a lower-redshift population than the simulations. As a result, the degeneracies associated with photometric redshifts described in chapter 4 where high-redshift sources can be confused for low-redshift sources are less of a problem. This is because erroneous inferences of sources being at high redshifts are *a priori* unlikely in a low-redshift sample.

It is also noteworthy that while the GMM approach in chapter 7 achieves a lower RMS scatter in both the simulated and GAMA blended source catalogue (Holwerda et al., 2015) tests, it fails to identify many of the sources in the latter test as being blended.

Since the GMM model is trained on real unblended sources, its flux-redshift distribution well represents the types of sources selected into the GAMA survey. One possible explanation of the results compared to chapter 6 is that the GMM model was able to constrain the redshifts by using strongly distinguishing features of the galaxy spectra, e.g. the Lyman and Balmer breaks, while failing to adequately describe the distribution of blended fluxes overall. This failure may be expected due to features of blended galaxies such as absorption due to dust that model does not account for. Such an effect would alter the shape of the spectrum without altering the position of strong break features, leading to the above behaviour. Since the Bayesian model selection procedure disfavours more complicated models without a sufficiently large increase in the likelihood, we would expect the above behaviour to result in a reasonably low scatter but poor identification of blended sources, as was observed.

The template-based model in Chapter 6 identifies more blended sources but also achieves a larger scatter. This could be explained by the template set characterising the flux distribution of GAMA galaxies less well than the empirically-trained model. If this were the case, we would expect that the scatter in the template-based model would be higher, as was observed. However, the additional freedom afforded by the blended model would not be as discouraged as in the GMM case by the Bayesian model comparison, since the single-constituent model being compared against is a poorer description of the data than the corresponding comparison in chapter 7. As a result, we would expect to see a higher scatter but better identification of blended sources than the empirical model.

The aim of the methods presented in this thesis is to generalise inferences of photometric redshifts to the case of blended sources, ultimately improving cosmological parameter inferences over neglecting these sources. However, it is difficult to estimate how cosmological parameter inferences are affected by neglecting blending. Photometric redshifts are only one aspect of the cosmological analysis pipeline that we would expect to be affected by unaccounted-for blending.

The priors in template based methods are generally useful in reducing the rate of catastrophic outliers (e.g., Benítez, 2000), but blending-related selection effects would make these priors less applicable to the survey population, increasing the rate of catastrophic outliers. These catastrophic outliers are able to induce systematic biases in inferences of cosmological parameters (e.g., Hearin et al., 2010). In addition, machine learning-based methods are particularly sensitive to the degree to which their training sets are representative of the population on which they are applied. Blended sources are likely to affect this, reducing the accuracy of these methods. The impact of such effects and how they interact with other effects of blended sources throughout the rest of the analysis pipeline would need to be tested through simulation of the full data-generating process.

We can test for the degree to which photometric redshifts are affected by neglecting blending by applying a standard photometric redshift method that does not account for blending to simulated blended data. To do this, we use simulated blended LSST-like data as described in chapter 7 and apply the GMM method described therein assuming the number of constituents $N = 1$ and $N = 2$. We then assign sources to tomographic bins of width $\Delta z = 0.1$ based on the point estimate of the redshift given by the mean of the posterior samples. Finally, we stack the redshift posteriors corresponding to each tomographic bin and obtain the mean redshift in each.

This mean redshift of sources in each tomographic bin is an important quantity for inferring cosmological parameters. For example, the dark-energy science goals of LSST require the error on this value to be $< 0.003(1 + z)$ with a goal of $< 0.005(1 + z)$ after the first year (The LSST Dark Energy Science Collaboration et al., 2018). Figure 9.1 shows the distribution of this error over tomographic bins for simulated blended data analysed assuming $N = 1$ and $N = 2$. We find the average of this error over all tomographic bins to be $0.226(1 + z)$ for the analysis that neglects blending, while the error reduces to $0.054(1 + z)$ when using the blended photometric redshift method.

While the error from the blended photometric redshift method is still larger than required for LSST, we emphasise that this is only simple test. In particular, stacking posteriors is a poor method for obtaining redshift distributions, as discussed in

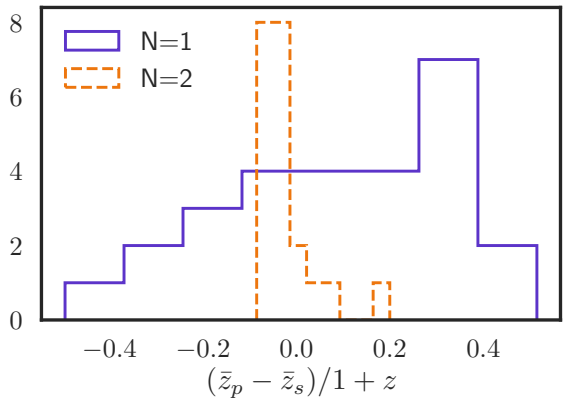

Figure 9.1: Distribution of error in the mean redshift in each tomographic bin, calculated from simulated LSST-like data. The solid purple line shows the results for a standard photometric redshift method that neglects blending, while the dashed orange line shows the results for a blended photometric redshift analysis.

chapter 4. The simulated observations are also only somewhat representative of future LSST observations as they do not simulate the observational pipeline in detail. A more complete understanding of the cosmological impact of blending would come from applying well-established methods for inferring redshift distributions, alongside the rest of the data analysis pipeline, on detailed simulated observations. Nevertheless, this test demonstrates that a photometric redshift method applied to blended data that does not account for this blending results in an increased error in cosmologically relevant metrics and will ultimately have an impact on the resulting cosmological parameter inferences.

Several additional avenues to extend the work in this thesis are possible. Firstly, since machine learning methods do not require a mathematical formulation of the forward model of the data, it is simple to train them for a variety of tasks. It would be worthwhile investigating how well a machine learning-based photometric redshift method trained with a training set of blended sources could recover vectors of redshifts. The single-constituent results presented in chapter 7 are also applicable for predicting a vector output. However, without the need to be able to construct blended posteriors from unblended training data, any machine learning methods could be used.

A potential problem for using blended training data is ensuring that the training set is sufficiently large and representative. Examples of all *pairs* blended training data would need to be available for training, increasing the size of the required training set. One possible solution to this problem would be to utilise multi-task learning, where a machine learning method such as a neural network is trained to perform multiple different predictions tasks. This could be applied to the blended photometric redshift

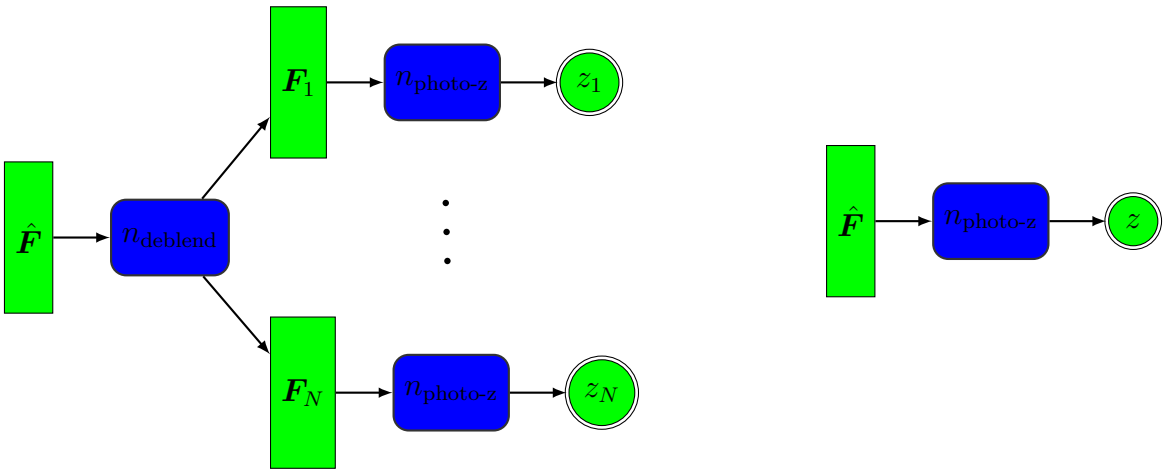

Figure 9.2: Figure showing the two modes of the possible multi-task blended photometric redshift neural network described in the text. The left figure shows the blended mode, where the network first predicts unblended fluxes, before using these vectors as input into multiple copies of the photo-z network with shared weights to predict the redshifts. The right figure shows the unblended mode, where only the photo-z network is used with unblended fluxes as input.

problem as follows.

A neural network model can be considered in two distinct sets of layers, as shown in Figure 9.2. The first set, labelled $n_{\text{photo-z}}$ is used to map fluxes of unblended sources to a single redshift. The second set of layers, labelled n_{deblend} is used to map blended fluxes onto multiple sets of unblended fluxes. The network can then be trained using standard backpropagation techniques using both blended and unblended training data. When unblended training data is used, only the $n_{\text{photo-z}}$ layers are backpropagated through. The $n_{\text{photo-z}}$ layers therefore learn how to predict photometric redshifts. However, when blended training data is used, the n_{deblend} layers first output unblended flux vectors that are then each used as input to one of several copies of the $n_{\text{photo-z}}$ layers with shared weights. Thus, by backpropagating through all layers, the deblending network can be trained to predict fluxes that the photo-z network maps to the correct redshift.

Another potential extension of this work is to using images as inputs. The work in this thesis shows a different approach to the typical deblending approach by inferring the quantities of interest from blended data directly. This is done by constructing a forward model of the blended data. If a forward model of galaxy images could be constructed, the same method could be applied to images themselves. This forward model could be a simple parametric model such as an Gaussian profile, convolved with the telescope beam. Generative machine learning models such as conditional variational autoencoders could also be used for this purpose, which may be more computationally

efficient and thus able to scale to the large datasets of future surveys.

A cosmological application of this method would be shape measurement, a necessary measurement for weak lensing surveys. By having both the brightness in each band of each constituent in a blended source and their corresponding shapes as inputs to the forward image model, these parameters could be fitted to blended images, providing joint distributions over the shapes and photometric redshifts of all constituents in a blended source.

Future cosmological galaxy surveys hold much promise for great increases in the precision of cosmological constraints, and an increased understanding of cosmology in general as a result. If these progressions are to come to fruition, much care will need to be taken to ensure we thoroughly understand and account for all uncertainties in these results. The very large datasets these surveys will produce promise to provide a significant challenge for existing statistical analyses, and are a scenario well-suited to machine learning methods. By combining these methods with existing statistical viewpoints to enable probabilistic approaches to machine learning, we can tackle both of these statistical and big data challenges, enabling the progress promised by a new generation of cosmological galaxy surveys.

Bibliography

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283 (2016).

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., Affeldt, C., Agathos, M., Agatsuma, K., Aggarwal, N., Aguiar, O. D., Aiello, L., Ain, A., Ajith, P., Allen, B., et al. Observation of gravitational waves from a binary black hole merger. *Phys. Rev. Lett.*, 116:061102 (2016). doi:10.1103/PhysRevLett.116.061102.

Abbott, B. P., Abbott, R., Abbott, T. D., Acernese, F., Ackley, K., Adams, C., Adams, T., Addesso, P., Adhikari, R. X., Adya, V. B., Affeldt, C., Agarwal, B., Agathos, M., Agatsuma, K., Aggarwal, N., Aguiar, O. D., Aiello, L., Ain, A., Ajith, P., Allen, B., et al. Tests of general relativity with gw170817. *Phys. Rev. Lett.*, 123:011102 (2019). doi:10.1103/PhysRevLett.123.011102.

Akrami, Y., Kallosh, R., Linde, A., and Vardanyan, V. The landscape, the swampland and the era of precision cosmology. *arXiv e-prints*, arXiv:1808.09440 (2018).

Alam, S., Ata, M., Bailey, S., Beutler, F., Bizyaev, D., Blazek, J. A., Bolton, A. S., Brownstein, J. R., Burden, A., Chuang, C.-H., Comparat, J., Cuesta, A. J., Dawson, K. S., Eisenstein, D. J., Escoffier, S., Gil-Marín, H., Grieb, J. N., Hand, N., Ho, S., Kinemuchi, K., et al. The clustering of galaxies in the completed SDSS-III Baryon

Oscillation Spectroscopic Survey: cosmological analysis of the DR12 galaxy sample. *MNRAS*, 470:2617–2652 (2017). doi:10.1093/mnras/stx721.

Alarcon, A., Sánchez, C., Bernstein, G. M., and Gaztañaga, E. Redshift inference from the combination of galaxy colors and clustering in a hierarchical Bayesian model – Application to realistic N-body simulations. *arXiv e-prints*, arXiv:1910.07127 (2019).

Albrecht, A., Bernstein, G., Cahn, R., Freedman, W. L., Hewitt, J., Hu, W., Huth, J., Kamionkowski, M., Kolb, E. W., Knox, L., Mather, J. C., Staggs, S., and Suntzeff, N. B. Report of the Dark Energy Task Force. *arXiv e-prints*, astro-ph/0609591 (2006).

Allison, R., Caucal, P., Calabrese, E., Dunkley, J., and Louis, T. Towards a cosmological neutrino mass detection. *Phys. Rev. D*, 92(12):123535 (2015). doi:10.1103/PhysRevD.92.123535.

Almosallam, I. A., Jarvis, M. J., and Roberts, S. J. GPZ: non-stationary sparse Gaussian processes for heteroscedastic uncertainty estimation in photometric redshifts. *MNRAS*, 462:726–739 (2016). doi:10.1093/mnras/stw1618.

Alonso, D., Ferreira, P. G., Jarvis, M. J., and Moodley, K. Calibrating photometric redshifts with intensity mapping observations. *Phys. Rev. D*, 96(4):043515 (2017). doi:10.1103/PhysRevD.96.043515.

Alsing, J., Charnock, T., Feeney, S., and Wand elt, B. Fast likelihood-free cosmology with neural density estimators and active learning. *MNRAS*, page 1888 (2019). doi:10.1093/mnras/stz1960.

Alsing, J., Heavens, A., and Jaffe, A. H. Cosmological parameters, shear maps and power spectra from CFHTLenS using Bayesian hierarchical inference. *MNRAS*, 466(3):3272–3292 (2017). doi:10.1093/mnras/stw3161.

Alsing, J., Heavens, A., Jaffe, A. H., Kiessling, A., Wandelt, B., and Hoffmann, T. Hierarchical cosmic shear power spectrum inference. *MNRAS*, 455(4):4452–4466 (2016). doi:10.1093/mnras/stv2501.

Amaro, V., Cavuoti, S., Brescia, M., Vellucci, C., Longo, G., Bilicki, M., de Jong, J. T., Tortora, C., Radovich, M., Napolitano, N. R., et al. Statistical analysis of probability density functions for photometric redshifts through the kids-eso-dr3 galaxies. *Monthly Notices of the Royal Astronomical Society*, 482(3):3116–3134 (2018). doi: 10.1093/mnras/sty2922.

Attias, H. A variational bayesian framework for graphical models. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, pages 209–215. MIT Press, Cambridge, MA, USA (1999).

Baldry, I. K., Alpaslan, M., Bauer, A. E., Bland-Hawthorn, J., Brough, S., Cluver, M. E., Croom, S. M., Davies, L. J. M., Driver, S. P., Gunawardhana, M. L. P., Holwerda, B. W., Hopkins, A. M., Kelvin, L. S., Liske, J., López-Sánchez, Á. R., Loveday, J., Norberg, P., Peacock, J., Robotham, A. S. G., and Taylor, E. N. Galaxy And Mass Assembly (GAMA): AUTOZ spectral redshift measurements, confidence and errors. *MNRAS*, 441:2440–2451 (2014). doi:10.1093/mnras/stu727.

Baldry, I. K., Liske, J., Brown, M. J. I., Robotham, A. S. G., Driver, S. P., Dunne, L., Alpaslan, M., Brough, S., Cluver, M. E., Eardley, E., Farrow, D. J., Heymans, C., Hildebrandt, H., Hopkins, A. M., Kelvin, L. S., Loveday, J., Moffett, A. J., Norberg, P., Owers, M. S., Taylor, E. N., et al. Galaxy And Mass Assembly (GAMA): the G02 field, Herschel-ATLAS target selection and Data Release 3. *ArXiv e-prints* (2017).

Bardeen, J. M., Bond, J., Kaiser, N., and Szalay, A. The statistics of peaks of gaussian random fields. *The Astrophysical Journal*, 304:15–61 (1986).

Bardeen, J. M., Steinhardt, P. J., and Turner, M. S. Spontaneous creation of almost scale-free density perturbations in an inflationary universe. *Physical Review D*, 28(4):679 (1983).

Bates, D. J., Tojeiro, R., Newman, J. A., Gonzalez-Perez, V., Comparat, J., Schneider, D. P., Lima, M., and Streblyanska, A. Mass functions, luminosity functions, and

completeness measurements from clustering redshifts. *MNRAS*, 486(3):3059–3077 (2019). doi:10.1093/mnras/stz997.

Bayarri, M. J. and Berger, J. O. The interplay of bayesian and frequentist analysis. *Statistical Science*, pages 58–80 (2004).

Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. Automatic differentiation in machine learning: a survey. *Journal of machine learning research*, 18(153) (2018).

Bayes, T. An essay towards solving a problem in the doctrine of chances. by the late rev. mr. bayes, frs communicated by mr. price, in a letter to john canton, amfr s. *Philosophical transactions of the Royal Society of London*, (53):370–418 (1763).

Beard, S. M., MacGillivray, H. T., and Thanisch, P. F. The Cosmos system for crowded-field analysis of digitized photographic plate scans. *MNRAS*, 247:311–321 (1990).

Beck, R., Lin, C.-A., Ishida, E. E. O., Gieseke, F., de Souza, R. S., Costa-Duarte, M. V., Hattab, M. W., and Krone-Martins, A. On the realistic validation of photometric redshifts. *MNRAS*, 468:4323–4339 (2017). doi:10.1093/mnras/stx687.

Benítez, N. Bayesian Photometric Redshift Estimation. *ApJ*, 536:571–583 (2000). doi:10.1086/308947.

Benítez, N., Moles, M., Aguerri, J. A. L., Alfaro, E., Broadhurst, T., Cabrera-Caño, J., Castand er, F. J., Cepa, J., Cerviño, M., Cristóbal-Hornillos, D., Fernández-Soto, A., González Delgado, R. M., Infante, L., Márquez, I., Martínez, V. J., Masegosa, J., Del Olmo, A., Perea, J., Prada, F., Quintana, J. M., et al. Optimal Filter Systems for Photometric Redshift Estimation. *ApJ*, 692(1):L5–L8 (2009). doi:10.1088/0004-637X/692/1/L5.

Berger, J. O., Bernardo, J. M., and Sun, D. The formal definition of reference priors. *arXiv e-prints*, arXiv:0904.0156 (2009).

Bernardo, J. M. Reference posterior distributions for bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2):113–128 (1979). doi: 10.1111/j.2517-6161.1979.tb01066.x.

Bernoulli, J. *Ars conjectandi*. Impensis Thurnisiorum, fratrum (1713).

Bertin, E. and Arnouts, S. SExtractor: Software for source extraction. *A&AS*, 117:393–404 (1996). doi:10.1051/aas:1996164.

Betancourt, M. Cruising the simplex: Hamiltonian Monte Carlo and the Dirichlet distribution. In Goyal, P., Giffin, A., Knuth, K. H., and Vrscay, E., editors, *American Institute of Physics Conference Series*, volume 1443 of *American Institute of Physics Conference Series*, pages 157–164 (2012). doi:10.1063/1.3703631.

Betancourt, M. Identifying bayesian mixture models. `https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html` (2017). Accessed: 2019-09-23.

Beutler, F., Blake, C., Colless, M., Jones, D. H., Staveley-Smith, L., Campbell, L., Parker, Q., Saunders, W., and Watson, F. The 6dF Galaxy Survey: baryon acoustic oscillations and the local Hubble constant. *MNRAS*, 416:3017–3032 (2011). doi: 10.1111/j.1365-2966.2011.19250.x.

Bilicki, M., Hoekstra, H., Brown, M. J. I., Amaro, V., Blake, C., Cavuoti, S., de Jong, J. T. A., Georgiou, C., Hildebrandt, H., Wolf, C., Amon, A., Brescia, M., Brough, S., Costa-Duarte, M. V., Erben, T., Glazebrook, K., Grado, A., Heymans, C., Jarrett, T., Joudaki, S., et al. Photometric redshifts for the Kilo-Degree Survey. Machine-learning analysis with artificial neural networks. *A&A*, 616:A69 (2018). doi:10.1051/0004-6361/201731942.

Bishop, C. M. Mixture density networks (1994).

Bishop, C. M. *Pattern recognition and machine learning*. springer (2006).

Blanton, M. R., Dalcanton, J., Eisenstein, D., Loveday, J., Strauss, M. A., SubbaRao, M., Weinberg, D. H., Anderson, John E., J., Annis, J., Bahcall, N. A., Bernardi, M., Brinkmann, J., Brunner, R. J., Burles, S., Carey, L., Castander, F. J., Connolly, A. J., Csabai, I., Doi, M., Finkbeiner, D., et al. The Luminosity Function of Galaxies in SDSS Commissioning Data. *AJ*, 121(5):2358–2380 (2001). doi:10.1086/320405.

Blei, D. M., Jordan, M. I., et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143 (2006).

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877 (2017).

Blei, D. M., Ng, A. Y., and Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022 (2003).

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. Weight Uncertainty in Neural Networks. *arXiv e-prints*, arXiv:1505.05424 (2015).

Bolzonella, M., Miralles, J.-M., and Pelló, R. Photometric redshifts based on standard SED fitting procedures. *A&A*, 363:476–492 (2000).

Bonnett, C., Troxel, M. A., Hartley, W., Amara, A., Leistedt, B., Becker, M. R., Bernstein, G. M., Bridle, S. L., Bruderer, C., Busha, M. T., Carrasco Kind, M., Childress, M. J., Castander, F. J., Chang, C., Crocce, M., Davis, T. M., Eifler, T. F., Frieman, J., Gangkofner, C., Gaztanaga, E., et al. Redshift distributions of galaxies in the Dark Energy Survey Science Verification shear catalogue and implications for weak lensing. *Phys. Rev. D*, 94(4):042005 (2016). doi:10.1103/PhysRevD.94.042005.

Boole, G. *The mathematical analysis of logic*. Philosophical Library (1847).

Bottou, L. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer (2010).

Bovy, J., Hogg, D. W., and Roweis, S. T. Extreme deconvolution: Inferring complete distribution functions from noisy, heterogeneous and incomplete observations. *Annals of Applied Statistics*, 5:1657–1677 (2011). doi:10.1214/10-AOAS439.

Bovy, J., Myers, A. D., Hennawi, J. F., Hogg, D. W., McMahon, R. G., Schiminovich, D., Sheldon, E. S., Brinkmann, J., Schneider, D. P., and Weaver, B. A. Photometric Redshifts and Quasar Probabilities from a Single, Data-driven Generative Model. *ApJ*, 749:41 (2012). doi:10.1088/0004-637X/749/1/41.

Brammer, G. B., van Dokkum, P. G., and Coppi, P. EAZY: A Fast, Public Photometric Redshift Code. *ApJ*, 686:1503–1513 (2008). doi:10.1086/591786.

Breiman, L. Random forests. *Machine learning*, 45(1):5–32 (2001).

Brewer, B. J., Pártay, L. B., and Csányi, G. Diffusive Nested Sampling. *arXiv e-prints*, arXiv:0912.2380 (2009).

Briggs, W. M. It is Time to Stop Teaching Frequentism to Non-statisticians. *arXiv e-prints*, arXiv:1201.2590 (2012).

Brimioulle, F., Lerchster, M., Seitz, S., Bender, R., and Snigula, J. Photometric redshifts for the CFHTLS-Wide. *ArXiv e-prints* (2008).

Buchdahl, H. A. Non-Linear Lagrangians and Cosmological Theory. *Monthly Notices of the Royal Astronomical Society*, 150(1):1–8 (1970). ISSN 0035-8711. doi:10.1093/mnras/150.1.1.

Buchs, R., Davis, C., Gruen, D., DeRose, J., Alarcon, A., Bernstein, G. M., Sánchez, C., Myles, J., Roodman, A., Allen, S., Amon, A., Choi, A., Masters, D. C., Miquel, R., Troxel, M. A., Wechsler, R. H., Abbott, T. M. C., Annis, J., Avila, S., Bechtol, K., et al. Phenotypic redshifts with self-organizing maps: A novel method to characterize redshift distributions of source galaxies for weak lensing. *MNRAS*, 489(1):820–841 (2019). doi:10.1093/mnras/stz2162.

Burke, C. J., Aleo, P. D., Chen, Y.-C., Liu, X., Peterson, J. R., Sembroski, G. H., and Yao-Yu Lin, J. Deblending and Classifying Astronomical Sources with Mask R-CNN Deep Learning. *arXiv e-prints*, arXiv:1908.02748 (2019).

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208 (1995).

Carliles, S., Budavári, T., Heinis, S., Priebe, C., and Szalay, A. S. Random Forests for Photometric Redshifts. *ApJ*, 712:511–515 (2010). doi:10.1088/0004-637X/712/1/511.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1) (2017).

Carpenter, B., Hoffman, M. D., Brubaker, M., Lee, D., Li, P., and Betancourt, M. The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164* (2015).

Carrasco Kind, M. and Brunner, R. J. TPZ: photometric redshift PDFs and ancillary information by using prediction trees and random forests. *MNRAS*, 432:1483–1501 (2013). doi:10.1093/mnras/stt574.

Carrasco Kind, M. and Brunner, R. J. Sparse representation of photometric redshift probability density functions: preparing for petascale astronomy. *MNRAS*, 441:3550–3561 (2014). doi:10.1093/mnras/stu827.

Carroll, S. Dark Energy and the Preposterous Universe. In *KITP: Colloquium Series*, page 2 (2002).

Chandrasekhar, S. The Maximum Mass of Ideal White Dwarfs. *ApJ*, 74:81 (1931). doi:10.1086/143324.

Chang, C., Jarvis, M., Jain, B., Kahn, S. M., Kirkby, D., Connolly, A., Krughoff, S., Peng, E.-H., and Peterson, J. R. The effective number density of galaxies for weak lensing measurements in the LSST project. *MNRAS*, 434:2121–2135 (2013). doi:10.1093/mnras/stt1156.

Chevallier, M. and Polarski, D. Accelerating Universes with Scaling Dark Matter. *International Journal of Modern Physics D*, 10:213–223 (2001). doi:10.1142/S0218271801000822.

Chisari, N. E., Richardson, M. L. A., Devriendt, J., Dubois, Y., Schneider, A., Le Brun, A. M. C., Beckmann, R. S., Peirani, S., Slyz, A., and Pichon, C. The impact of baryons on the matter power spectrum from the Horizon-AGN cosmological hydrodynamical simulation. *MNRAS*, 480(3):3962–3977 (2018). doi:10.1093/mnras/sty2093.

Choromanska, A., Henaff, M., Mathieu, M., Ben Arous, G., and LeCun, Y. The Loss Surfaces of Multilayer Networks. *arXiv e-prints*, arXiv:1412.0233 (2014).

Choudhury, S. R. and Choubey, S. Updated bounds on sum of neutrino masses in various cosmological scenarios. *Journal of Cosmology and Astro-Particle Physics*, 2018(9):017 (2018). doi:10.1088/1475-7516/2018/09/017.

Clements, D. L. An introduction to the Planck mission. *Contemporary Physics*, 58(4):331–348 (2017). doi:10.1080/00107514.2017.1362139.

Clifton, T., Ferreira, P. G., Padilla, A., and Skordis, C. Modified gravity and cosmology. *Physics Reports*, 513(1):1 – 189 (2012). ISSN 0370-1573. doi:https://doi.org/10.1016/j.physrep.2012.01.001. Modified Gravity and Cosmology.

Coe, D., Benítez, N., Sánchez, S. F., Jee, M., Bouwens, R., and Ford, H. Galaxies in the Hubble Ultra Deep Field. I. Detection, Multiband Photometry, Photometric Redshifts, and Morphology. *AJ*, 132:926–959 (2006). doi:10.1086/505530.

Coleman, G. D., Wu, C.-C., and Weedman, D. W. Colors and magnitudes predicted for high redshift galaxies. *ApJS*, 43:393–416 (1980). doi:10.1086/190674.

Collister, A. A. and Lahav, O. ANNz: Estimating Photometric Redshifts Using Artificial Neural Networks. *PASP*, 116:345–351 (2004). doi:10.1086/383254.

Connolly, A. J., Csabai, I., Szalay, A. S., Koo, D. C., Kron, R. G., and Munn, J. A. Slicing Through Multicolor Space: Galaxy Redshifts from Broadband Photometry. *AJ*, 110:2655 (1995). doi:10.1086/117720.

Cooke, R. J., Pettini, M., Nollett, K. M., and Jorgenson, R. The Primordial Deuterium Abundance of the Most Metal-poor Damped Lyman-α System. *ApJ*, 830:148 (2016). doi:10.3847/0004-637X/830/2/148.

Cortês, M. V. *The Old and New Universe in the era of precision cosmology*. Ph.D. thesis, University of Sussex (2010).

Cox, R. T. Probability, frequency and reasonable expectation. *American journal of physics*, 14(1):1–13 (1946).

Dang, K.-D., Quiroz, M., Kohn, R., Tran, M.-N., and Villani, M. Hamiltonian Monte Carlo with Energy Conserving Subsampling. *arXiv e-prints*, arXiv:1708.00955 (2017).

Dark Energy Survey Collaboration, Abbott, T., Abdalla, F. B., Aleksić, J., Allam, S., Amara, A., Bacon, D., Balbinot, E., Banerji, M., Bechtol, K., Benoit-Lévy, A., Bernstein, G. M., Bertin, E., Blazek, J., Bonnett, C., Bridle, S., Brooks, D., Brunner, R. J., Buckley-Geer, E., Burke, D. L., et al. The Dark Energy Survey: more than dark energy - an overview. *MNRAS*, 460:1270–1299 (2016). doi:10.1093/mnras/stw641.

Dark Energy Survey Collaboration, Abbott, T. M. C., Abdalla, F. B., Annis, J., Bechtol, K., Blazek, J., Benson, B. A., Bernstein, R. A., Bernstein, G. M., Bertin, E., Brooks, D., Burke, D. L., Carnero Rosell, A., Carrasco Kind, M., Carretero, J., Castander, F. J., Chang, C. L., Crawford, T. M., Cunha, C. E., D'Andrea, C. B., et al. Dark Energy Survey Year 1 Results: A Precise H_0 Estimate from DES Y1, BAO, and D/H Data. *MNRAS*, 480(3):3879–3888 (2018). doi:10.1093/mnras/sty1939.

David, H. A. and Nagaraja, H. N. *Order Statistics*. American Cancer Society (2006). ISBN 9780471667193. doi:10.1002/0471667196.ess6023.pub2.

Dawson, W. and Schneider, M. Complementarity of lsst and wfirst: Regarding object blending. Technical report, Lawrence Livermore National Laboratory (LLNL), Livermore, CA (2014).

Dawson, W. A., Schneider, M. D., Tyson, J. A., and Jee, M. J. The Ellipticity Distribution of Ambiguously Blended Objects. *ApJ*, 816:11 (2016). doi:10.3847/0004-637X/816/1/11.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22 (1977). doi:10.1111/j.2517-6161.1977.tb01600.x.

Devroye, L. *Non-Uniform Random Variate Generation*. Springer-Verlag New York (1986). doi:10.1007/978-1-4613-8643-8.

Dewdney, P., Hall, P., Schillizzi, R., and Lazio, J. The square kilometre array. *Proceedings of the Institute of Electrical and Electronics Engineers IEEE*, 97(8):1482–1496 (2009).

Dickey, J. M. The weighted likelihood ratio, linear hypotheses on normal location parameters. *The Annals of Mathematical Statistics*, pages 204–223 (1971).

D'Isanto, A. and Polsterer, K. L. Photometric redshift estimation via deep learning. Generalized and pre-classification-less, image based, fully probabilistic redshifts. *A&A*, 609:A111 (2018). doi:10.1051/0004-6361/201731326.

Dolgov, A. D. Neutrinos in cosmology. *Physics Reports*, 370(4):333–535 (2002).

Dozat, T. Incorporating nesterov momentum into adam (2016).

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. Hybrid monte carlo. *Physics letters B*, 195(2):216–222 (1987).

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159 (2011).

Duncan, K. J., Jarvis, M. J., Brown, M. J., and Röttgering, H. J. Photometric redshifts for the next generation of deep radio continuum surveys–ii. gaussian processes and hybrid estimates. *Monthly Notices of the Royal Astronomical Society*, 477(4):5177–5190 (2018). doi:10.1093/mnras/sty940.

Duncan, K. J., Jarvis, M. J., Brown, M. J. I., and Röttgering, H. J. A. Photometric redshifts for the next generation of deep radio continuum surveys - II. Gaussian processes and hybrid estimates. *MNRAS*, 477:5177–5190 (2018). doi:10.1093/mnras/sty940.

Dyson, F. A meeting with enrico fermi. *Nature*, 427(6972):297 (2004). doi:10.1038/427297a.

Dyson, F. W., Eddington, A. S., and Davidson, C. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of may 29, 1919. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 220(571-581):291–333 (1920).

Edge, A., Sutherland, W., Kuijken, K., Driver, S., McMahon, R., Eales, S., and Emerson, J. P. The VISTA Kilo-degree Infrared Galaxy (VIKING) Survey: Bridging the Gap between Low and High Redshift. *The Messenger*, 154:32–34 (2013).

Efron, B. Maximum likelihood and decision theory. *The annals of Statistics*, pages 340–356 (1982).

Einstein, A. Die Feldgleichungen der Gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin), Seite 844-847.* (1915).

Einstein, A. Die grundlage der allgemeinen relativitätstheorie. *Annalen der Physik*, 354(7):769–822 (1916).

Einstein, A. Kosmologische Betrachtungen zur allgemeinen Relativitätstheorie. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin*, pages 142–152 (1917).

Eisenstein, D. J., Zehavi, I., Hogg, D. W., Scoccimarro, R., Blanton, M. R., Nichol, R. C., Scranton, R., Seo, H.-J., Tegmark, M., Zheng, Z., Anderson, S. F., Annis, J., Bahcall, N., Brinkmann, J., Burles, S., Castand er, F. J., Connolly, A., Csabai, I., Doi, M., Fukugita, M., et al. Detection of the Baryon Acoustic Peak in the Large-Scale Correlation Function of SDSS Luminous Red Galaxies. *ApJ*, 633(2):560–574 (2005). doi:10.1086/466512.

Fabian, A. Observational evidence of active galactic nuclei feedback. *Annual Review of Astronomy and Astrophysics*, 50:455–489 (2012).

Feeney, S. M., Mortlock, D. J., and Dalmasso, N. Clarifying the Hubble constant tension with a Bayesian hierarchical model of the local distance ladder. *MNRAS*, 476(3):3861–3882 (2018). doi:10.1093/mnras/sty418.

Feldmann, R., Carollo, C. M., Porciani, C., Lilly, S. J., Capak, P., Taniguchi, Y., Le Fèvre, O., Renzini, A., Scoville, N., Ajiki, M., Aussel, H., Contini, T., McCracken, H., Mobasher, B., Murayama, T., Sanders, D., Sasaki, S., Scarlata, C., Scodeggio, M., Shioya, Y., et al. The Zurich Extragalactic Bayesian Redshift Analyzer and its first application: COSMOS. *MNRAS*, 372:565–577 (2006). doi:10.1111/j.1365-2966.2006.10930.x.

Feroz, F., Hobson, M. P., and Bridges, M. MULTINEST: an efficient and robust Bayesian inference tool for cosmology and particle physics. *MNRAS*, 398:1601–1614 (2009). doi:10.1111/j.1365-2966.2009.14548.x.

Firth, A. E., Lahav, O., and Somerville, R. S. Estimating photometric redshifts with artificial neural networks. *Monthly Notices of the Royal Astronomical Society*, 339(4):1195–1202 (2003).

Fixsen, D. J. The Temperature of the Cosmic Microwave Background. *ApJ*, 707:916–920 (2009). doi:10.1088/0004-637X/707/2/916.

Fixsen, D. J., Cheng, E. S., Cottingham, D. A., Eplee, Jr., R. E., Isaacman, R. B., Mather, J. C., Meyer, S. S., Noerdlinger, P. D., Shafer, R. A., Weiss, R., Wright, E. L., Bennett, C. L., Boggess, N. W., Kelsall, T., Moseley, S. H., Silverberg, R. F., Smoot, G. F., and Wilkinson, D. T. Cosmic microwave background dipole spectrum measured by the COBE FIRAS instrument. *ApJ*, 420:445–449 (1994). doi:10.1086/173575.

Foreman-Mackey, D. corner.py: Scatterplot matrices in python. *The Journal of Open Source Software*, 24 (2016). doi:10.21105/joss.00024.

Foreman-Mackey, D., Hogg, D. W., Lang, D., and Goodman, J. emcee: The MCMC Hammer. *PASP*, 125:306 (2013). doi:10.1086/670067.

Fragoso, T. M., Bertoli, W., and Louzada, F. Bayesian model averaging: A systematic review and conceptual classification. *International Statistical Review*, 86(1):1–28 (2018). doi:10.1111/insr.12243.

Fraley, C. and Raftery, A. E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The computer journal*, 41(8):578–588 (1998).

Friedmann, A. Über die krümmung des raumes. *Zeitschrift für Physik A Hadrons and Nuclei*, 10(1):377–386 (1922).

Fukugita, M., Ichikawa, T., Gunn, J. E., Doi, M., Shimasaku, K., and Schneider, D. P. The Sloan Digital Sky Survey Photometric System. *AJ*, 111:1748 (1996). doi:10.1086/117915.

Fukugita, M., Okamura, S., et al. Automated software for surface photometry of galaxies. *The Astrophysical Journal Supplement Series*, 97:59–75 (1995).

Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., Vallenari, A., Babusiaux, C., Bailer-Jones, C. A. L., Bastian, U., Biermann, M., Evans, D. W.,

Eyer, L., Jansen, F., Jordi, C., Klioner, S. A., Lammers, U., Lindegren, L., Luri, X., Mignard, F., Milligan, D. J., Panem, C., et al. The gaia mission. *A&A*, 595:A1 (2016). doi:10.1051/0004-6361/201629272.

Gamow, G. The evolutionary universe. *Scientific American*, 195(3):136–156 (1956).

Gatti, M., Vielzeuf, P., Davis, C., Cawthon, R., Rau, M. M., DeRose, J., De Vicente, J., Alarcon, A., Rozo, E., Gaztanaga, E., Hoyle, B., Miquel, R., Bernstein, G. M., Bonnett, C., Carnero Rosell, A., Castander, F. J., Chang, C., da Costa, L. N., Gruen, D., Gschwend, J., et al. Dark Energy Survey Year 1 results: cross-correlation redshifts - methods and systematics characterization. *MNRAS*, 477:1664–1682 (2018). doi:10.1093/mnras/sty466.

Gelman, A. Prior distribution. *Encyclopedia of environmetrics*, 4 (2006).

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. *Bayesian data analysis*. Chapman and Hall/CRC (2013).

Geman, S. and Geman, D. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6(6):721–741 (1984). ISSN 0162-8828. doi:10.1109/TPAMI.1984.4767596.

Gerbino, M. Neutrino mass scale in the era of precision cosmology. In *Journal of Physics: Conference Series*, volume 566, page 012003. IOP Publishing (2014).

Gerdes, D. W., Sypniewski, A. J., McKay, T. A., Hao, J., Weis, M. R., Wechsler, R. H., and Busha, M. T. ArborZ: Photometric Redshifts Using Boosted Decision Trees. *ApJ*, 715(2):823–832 (2010). doi:10.1088/0004-637X/715/2/823.

Ghahramani, Z. and Beal, M. J. Propagation algorithms for variational bayesian learning. In *Advances in neural information processing systems*, pages 507–513 (2001).

Gilks, W. R. and Wild, P. Adaptive rejection sampling for gibbs sampling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 41(2):337–348 (1992).

Girolami, M. and Calderhead, B. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214 (2011).

Gomes, Z., Jarvis, M. J., Almosallam, I. A., and Roberts, S. J. Improving photometric redshift estimation using gpz: size information, post processing, and improved photometry. *Monthly Notices of the Royal Astronomical Society*, 475(1):331–342 (2017). doi:10.1093/mnras/stx3187.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, arXiv:1412.6572 (2014).

Graham, M. L., Connolly, A. J., Ivezić, Ž., Schmidt, S. J., Jones, R. L., Jurić, M., Daniel, S. F., and Yoachim, P. Photometric Redshifts with the LSST: Evaluating Survey Observing Strategies. *AJ*, 155(1):1 (2018). doi:10.3847/1538-3881/aa99d4.

Graves, A. Practical variational inference for neural networks. In Shawe-Taylor, J., Zemel, R. S., Bartlett, P. L., Pereira, F., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 24*, pages 2348–2356. Curran Associates, Inc. (2011).

Greene, T. P., Chu, L., Egami, E., Hodapp, K. W., Kelly, D. M., Leisenring, J., Rieke, M., Robberto, M., Schlawin, E., and Stansberry, J. Slitless spectroscopy with the James Webb Space Telescope Near-Infrared Camera (JWST NIRCam). In *Proc. SPIE*, volume 9904 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, page 99040E (2016). doi:10.1117/12.2231347.

Gschwend, J., Rossel, A. C., Ogando, R. L. C., Neto, A. F., Maia, M. A. G., da Costa, L. N., Lima, M., Pellegrini, P., Campisano, R., Singulani, C., Adean, C., Benoist, C., Aguena, M., Carrasco Kind, M., Davis, T. M., de Vicente, J., Hartley, W. G., Hoyle, B., Palmese, A., Sadeh, I., et al. DES science portal: Computing photometric redshifts. *Astronomy and Computing*, 25:58–80 (2018). doi:10.1016/j.ascom.2018.08.008.

Guy, J., Astier, P., Baumont, S., Hardin, D., Pain, R., Regnault, N., Basa, S., Carlberg, R. G., Conley, A., Fabbro, S., Fouchez, D., Hook, I. M., Howell, D. A., Perrett, K., Pritchet, C. J., Rich, J., Sullivan, M., Antilogus, P., Aubourg, E., Bazin, G., et al. SALT2: using distant supernovae to improve the use of type Ia supernovae as distance indicators. *A&A*, 466(1):11–21 (2007). doi:10.1051/0004-6361:20066930.

Gwyn, S. D. J. and Hartwick, F. D. A. The Redshift Distribution and Luminosity Functions of Galaxies in the Hubble Deep Field. *ApJ*, 468:L77 (1996). doi:10.1086/310237.

Hamuy, M., Phillips, M. M., Suntzeff, N. B., Schommer, R. A., Maza, J., and Aviles, R. The Hubble Diagram of the Calan/Tololo Type IA Supernovae and the Value of HO. *AJ*, 112:2398 (1996). doi:10.1086/118191.

Handley, W. J., Hobson, M. P., and Lasenby, A. N. POLYCHORD: next-generation nested sampling. *MNRAS*, 453(4):4384–4398 (2015). doi:10.1093/mnras/stv1911.

Harnois-Déraps, J., van Waerbeke, L., Viola, M., and Heymans, C. Baryons, neutrinos, feedback and weak gravitational lensing. *Monthly Notices of the Royal Astronomical Society*, 450(2):1212–1223 (2015).

Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, 27(2):83–85 (2005).

Hastings, W. K. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109 (1970). ISSN 0006-3444. doi:10.1093/biomet/57.1.97.

Hearin, A. P., Zentner, A. R., Ma, Z., and Huterer, D. A General Study of the Influence of Catastrophic Photometric Redshift Errors on Cosmology with Cosmic Shear Tomography. *ApJ*, 720(2):1351–1369 (2010). doi:10.1088/0004-637X/720/2/1351.

Heavens, A. 3D weak lensing. *MNRAS*, 343(4):1327–1334 (2003). doi:10.1046/j.1365-8711.2003.06780.x.

Heavens, A. F. and Sellentin, E. Objective bayesian analysis of neutrino masses and hierarchy. *Journal of Cosmology and Astroparticle Physics*, 2018(04):047 (2018). doi:10.1088/1475-7516/2018/04/047.

Heitmann, K., Bingham, D., Lawrence, E., Bergner, S., Habib, S., Higdon, D., Pope, A., Biswas, R., Finkel, H., Frontiere, N., and Bhattacharya, S. The Mira-Titan Universe: Precision Predictions for Dark Energy Surveys. *ApJ*, 820(2):108 (2016). doi:10.3847/0004-637X/820/2/108.

Heymans, C., Van Waerbeke, L., Miller, L., Erben, T., Hildebrandt, H., Hoekstra, H., Kitching, T. D., Mellier, Y., Simon, P., Bonnett, C., Coupon, J., Fu, L., Harnois Déraps, J., Hudson, M. J., Kilbinger, M., Kuijken, K., Rowe, B., Schrabback, T., Semboloni, E., van Uitert, E., et al. CFHTLenS: the Canada-France-Hawaii Telescope Lensing Survey. *MNRAS*, 427:146–166 (2012). doi:10.1111/j.1365-2966.2012.21952.x.

Higson, E., Handley, W., Hobson, M., and Lasenby, A. Dynamic nested sampling: an improved algorithm for parameter estimation and evidence calculation. *arXiv e-prints*, arXiv:1704.03459 (2017).

Hildebrandt, H., Arnouts, S., Capak, P., Moustakas, L. A., Wolf, C., Abdalla, F. B., Assef, R. J., Banerji, M., Benítez, N., Brammer, G. B., Budavári, T., Carliles, S., Coe, D., Dahlen, T., Feldmann, R., Gerdes, D., Gillis, B., Ilbert, O., Kotulla, R., Lahav, O., et al. PHAT: PHoto-z Accuracy Testing. *A&A*, 523:A31 (2010). doi:10.1051/0004-6361/201014885.

Hildebrandt, H., Viola, M., Heymans, C., Joudaki, S., Kuijken, K., Blake, C., Erben, T., Joachimi, B., Klaes, D., Miller, L., Morrison, C. B., Nakajima, R., Verdoes Kleijn, G., Amon, A., Choi, A., Covone, G., de Jong, J. T. A., Dvornik, A., Fenech Conti, I., Grado, A., et al. KiDS-450: cosmological parameter constraints from

tomographic weak gravitational lensing. *MNRAS*, 465:1454–1498 (2017). doi:10.1093/mnras/stw2805.

Hill, D. T., Kelvin, L. S., Driver, S. P., Robotham, A. S. G., Cameron, E., Cross, N., Andrae, E., Baldry, I. K., Bamford, S. P., Bland-Hawthorn, J., Brough, S., Conselice, C. J., Dye, S., Hopkins, A. M., Liske, J., Loveday, J., Norberg, P., Peacock, J. A., Croom, S. M., Frenk, C. S., et al. Galaxy and Mass Assembly: FUV, NUV, ugrizYJHK Petrosian, Kron and Sérsic photometry. *MNRAS*, 412:765–799 (2011). doi:10.1111/j.1365-2966.2010.17950.x.

Hivon, E., Górski, K. M., Netterfield, C. B., Crill, B. P., Prunet, S., and Hansen, F. MASTER of the Cosmic Microwave Background Anisotropy Power Spectrum: A Fast Method for Statistical Analysis of Large and Complex Cosmic Microwave Background Data Sets. *ApJ*, 567(1):2–17 (2002). doi:10.1086/338126.

Hobson, M. P., Efstathiou, G. P., and Lasenby, A. N. *General relativity: an introduction for physicists*. Cambridge University Press (2006).

Hoekstra, H., Mellier, Y., van Waerbeke, L., Semboloni, E., Fu, L., Hudson, M. J., Parker, L. C., Tereno, I., and Benabed, K. First cosmic shear results from the canada-france-hawaii telescope wide synoptic legacy survey. *The Astrophysical Journal*, 647(1):116–127 (2006). doi:10.1086/503249.

Hoekstra, H., Viola, M., and Herbonnet, R. A study of the sensitivity of shape measurements to the input parameters of weak-lensing image simulations. *MNRAS*, 468:3295–3311 (2017). doi:10.1093/mnras/stx724.

Hoerl, A. E. and Kennard, R. W. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67 (1970). doi:10.1080/00401706.1970.10488634.

Hoffman, M. D. and Gelman, A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *arXiv e-prints*, arXiv:1111.4246 (2011).

Hogg, D. W. Distance measures in cosmology. *ArXiv e-prints* (1999).

Holwerda, B. W., Baldry, I. K., Alpaslan, M., Bauer, A., Bland-Hawthorn, J., Brough, S., Brown, M. J. I., Cluver, M. E., Conselice, C., Driver, S. P., Hopkins, A. M., Jones, D. H., López-Sánchez, Á. R., Loveday, J., Meyer, M. J., and Moffett, A. Galaxy And Mass Assembly (GAMA) blended spectra catalogue: strong galaxy-galaxy lens and occulting galaxy pair candidates. *MNRAS*, 449:4277–4287 (2015). doi:10.1093/mnras/stv589.

Horndeski, G. W. Second-order scalar-tensor field equations in a four-dimensional space. *Int. J. Theor. Phys.*, 10:363–384 (1974). doi:10.1007/BF01807638.

Hoyer, P. O. Non-negative matrix factorization with sparseness constraints. *Journal of machine learning research*, 5(Nov):1457–1469 (2004).

Hoyle, B., Gruen, D., Bernstein, G. M., Rau, M. M., De Vicente, J., Hartley, W. G., Gaztanaga, E., DeRose, J., Troxel, M. A., Davis, C., Alarcon, A., MacCrann, N., Prat, J., Sánchez, C., Sheldon, E., Wechsler, R. H., Asorey, J., Becker, M. R., Bonnett, C., Carnero Rosell, A., et al. Dark Energy Survey Year 1 Results: redshift distributions of the weak-lensing source galaxies. *MNRAS*, 478(1):592–610 (2018). doi:10.1093/mnras/sty957.

Hu, W. Power Spectrum Tomography with Weak Lensing. *ApJ*, 522:L21–L24 (1999). doi:10.1086/312210.

Hubble, E. A Relation between Distance and Radial Velocity among Extra-Galactic Nebulae. *Proceedings of the National Academy of Science*, 15(3):168–173 (1929). doi:10.1073/pnas.15.3.168.

Hulse, R. A. and Taylor, J. H. Discovery of a pulsar in a binary system. *ApJ*, 195:L51–L53 (1975). doi:10.1086/181708.

Huterer, D., Takada, M., Bernstein, G., and Jain, B. Systematic errors in future weak-lensing surveys: requirements and prospects for self-calibration. *MNRAS*, 366:101–114 (2006). doi:10.1111/j.1365-2966.2005.09782.x.

Ilbert, O., Arnouts, S., McCracken, H. J., Bolzonella, M., Bertin, E., Le Fèvre, O., Mellier, Y., Zamorani, G., Pellò, R., Iovino, A., Tresse, L., Le Brun, V., Bottini, D., Garilli, B., Maccagni, D., Picat, J. P., Scaramella, R., Scodeggio, M., Vettolani, G., Zanichelli, A., et al. Accurate photometric redshifts for the CFHT legacy survey calibrated using the VIMOS VLT deep survey. *A&A*, 457:841–856 (2006). doi: 10.1051/0004-6361:20065138.

Ivezić, Ž., Kahn, S. M., Tyson, J. A., Abel, B., Acosta, E., Allsman, R., Alonso, D., AlSayyad, Y., Anderson, S. F., Andrew, J., Angel, J. R. P., Angeli, G. Z., Ansari, R., Antilogus, P., Araujo, C., Armstrong, R., Arndt, K. T., Astier, P., Aubourg, É., Auza, N., et al. LSST: From Science Drivers to Reference Design and Anticipated Data Products. *ApJ*, 873(2):111 (2019). doi:10.3847/1538-4357/ab042c.

Jain, A. K. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666 (2010).

James, G., Witten, D., Hastie, T., and Tibshirani, R. *An introduction to statistical learning*, volume 112. Springer (2013).

Jasche, J., Leclercq, F., and Wandelt, B. D. Past and present cosmic structure in the SDSS DR7 main sample. *J. Cosmology Astropart. Phys.*, 2015(1):036 (2015). doi:10.1088/1475-7516/2015/01/036.

Jaynes, E. T. *Probability theory: The logic of science*. Cambridge university press (2003).

Jaynes, E. T. and Kempthorne, O. Confidence intervals vs bayesian intervals. In *Foundations of probability theory, statistical inference, and statistical theories of science*, pages 175–257. Springer (1976).

Jefferys, W. H. and Berger, J. O. Ockham's razor and bayesian analysis. *American Scientist*, 80(1):64–72 (1992).

Jeffreys, H. *The Theory of Probability*. The Clarendon Press, Oxford (1939).

Jew, L. and Grumitt, R. The spectral index of polarized diffuse Galactic emission between 30 and 44 GHz. *arXiv e-prints*, arXiv:1907.11426 (2019).

Joachimi, B., Cacciato, M., Kitching, T. D., Leonard, A., Mandelbaum, R., Schäfer, B. M., Sifón, C., Hoekstra, H., Kiessling, A., Kirk, D., et al. Galaxy alignments: An overview. *Space Science Reviews*, 193(1-4):1–65 (2015).

Jones, D. M. and Heavens, A. F. Bayesian photometric redshifts of blended sources. *MNRAS*, 483:2487–2505 (2019a). doi:10.1093/mnras/sty3279.

Jones, D. M. and Heavens, A. F. Gaussian Mixture Models for Blended Photometric Redshifts. *Monthly Notices of the Royal Astronomical Society* (2019b). ISSN 0035-8711. doi:10.1093/mnras/stz2687. Stz2687.

Joseph, R., Courbin, F., and Starck, J.-L. Multi-band morpho-Spectral Component Analysis Deblending Tool (MuSCADeT): Deblending colourful objects. *A&A*, 589:A2 (2016). doi:10.1051/0004-6361/201527923.

Joudaki, S., Blake, C., Johnson, A., Amon, A., Asgari, M., Choi, A., Erben, T., Glazebrook, K., Harnois-Déraps, J., Heymans, C., Hildebrandt, H., Hoekstra, H., Klaes, D., Kuijken, K., Lidman, C., Mead, A., Miller, L., Parkinson, D., Poole, G. B., Schneider, P., et al. KiDS-450 + 2dFLenS: Cosmological parameter constraints from weak gravitational lensing tomography and overlapping redshift-space galaxy clustering. *MNRAS*, 474:4894–4924 (2018). doi:10.1093/mnras/stx2820.

Joudaki, S., Hildebrandt, H., Traykova, D., Chisari, N. E., Heymans, C., Kannawadi, A., Kuijken, K., Wright, A. H., Asgari, M., Erben, T., Hoekstra, H., Joachimi, B., Miller, L., Tröster, T., and van den Busch, J. L. KiDS+VIKING-450 and DES-Y1 combined: Cosmology with cosmic shear. *arXiv e-prints*, arXiv:1906.09262 (2019).

Kass, R. E. and Raftery, A. E. Bayes factors. *Journal of the american statistical association*, 90(430):773–795 (1995).

Kimura, M., Maihara, T., Iwamuro, F., Akiyama, M., Tamura, N., Dalton, G. B., Takato, N., Tait, P., Ohta, K., Eto, S., et al. Fibre multi-object spectrograph

(fmos) for the subaru telescope. *Publications of the Astronomical Society of Japan*, 62(5):1135–1147 (2010).

King, S. Neutrino mass. *Contemporary Physics*, 48(4):195–211 (2007).

Kingma, D. P. and Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980 (2014).

Kinney, A. L., Calzetti, D., Bohlin, R. C., McQuade, K., Storchi-Bergmann, T., and Schmitt, H. R. Template Ultraviolet to Near-Infrared Spectra of Star-forming Galaxies and Their Application to K-Corrections. *ApJ*, 467:38 (1996). doi:10.1086/177583.

Kitzbichler, M. G. and White, S. D. M. The high-redshift galaxy population in hierarchical galaxy formation models. *MNRAS*, 376(1):2–12 (2007). ISSN 0035-8711. doi:10.1111/j.1365-2966.2007.11458.x.

Knoetig, M. L. Signal discovery, limits, and uncertainties with sparse on/off measurements: an objective bayesian analysis. *The Astrophysical Journal*, 790(2):106 (2014). doi:10.1088/0004-637X/790/2/106.

Kobayashi, T. Horndeski theory and beyond: a review. *Reports on Progress in Physics*, 82(8):086901 (2019). doi:10.1088/1361-6633/ab2429.

Köhlinger, F., Viola, M., Joachimi, B., Hoekstra, H., van Uitert, E., Hildebrandt, H., Choi, A., Erben, T., Heymans, C., Joudaki, S., Klaes, D., Kuijken, K., Merten, J., Miller, L., Schneider, P., and Valentijn, E. A. KiDS-450: the tomographic weak lensing power spectrum and constraints on cosmological parameters. *MNRAS*, 471(4):4412–4435 (2017). doi:10.1093/mnras/stx1820.

Kokoska, S. and Zwillinger, D. *CRC standard probability and statistics tables and formulae*. Crc Press (2000).

Kolmogorov, A. N. *Foundations of the theory of probability: Second English Edition*. Courier Dover Publications (1933).

Koyama, K. Cosmological tests of modified gravity. *Reports on Progress in Physics*, 79(4):046902 (2016). doi:10.1088/0034-4885/79/4/046902.

Krause, E., Eifler, T. F., Zuntz, J., Friedrich, O., Troxel, M. A., Dodelson, S., Blazek, J., Secco, L. F., MacCrann, N., Baxter, E., Chang, C., Chen, N., Crocce, M., DeRose, J., Ferte, A., Kokron, N., Lacasa, F., Miranda, V., Omori, Y., Porredon, A., et al. Dark Energy Survey Year 1 Results: Multi-Probe Methodology and Simulated Likelihood Analyses. *arXiv e-prints*, arXiv:1706.09359 (2017).

Kremer, J., Gieseke, F., Steenstrup Pedersen, K., and Igel, C. Nearest neighbor density ratio estimation for large-scale applications in astronomy. *Astronomy and Computing*, 12:67–72 (2015).

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. M. Automatic Variational Inference in Stan. *arXiv e-prints*, arXiv:1506.03431 (2015).

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. Automatic Differentiation Variational Inference. *arXiv e-prints*, arXiv:1603.00788 (2016).

Kullback, S. and Leibler, R. A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86 (1951). doi:10.1214/aoms/1177729694.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv e-prints*, arXiv:1607.02533 (2016).

Laplace, P. S. *Théorie analytique des probabilités*. Courcier (1820).

Laplace, P.-S. *Essai philosophique sur les probabilitiés...* H. Remy (1829).

Laureijs, R., Amiaux, J., Arduini, S., Augueres, J. ., Brinchmann, J., Cole, R., Cropper, M., Dabin, C., Duvet, L., Ealet, A., and et al. Euclid Definition Study Report. *ArXiv e-prints* (2011).

Leclercq, F. Bayesian optimization for likelihood-free cosmological inference. *Phys. Rev. D*, 98(6):063511 (2018). doi:10.1103/PhysRevD.98.063511.

LeCun, Y. A., Bottou, L., Orr, G. B., and Müller, K.-R. Efficient backprop. In *Neural networks: Tricks of the trade*, pages 9–48. Springer (2012).

Leistedt, B. and Hogg, D. W. Data-driven, interpretable photometric redshifts trained on heterogeneous and unrepresentative data. *The Astrophysical Journal*, 838(1):5 (2017). doi:10.3847/1538-4357/aa6332.

Leistedt, B. and Hogg, D. W. Hierarchical Probabilistic Inference of the Color-Magnitude Diagram and Shrinkage of Stellar Distance Uncertainties. *AJ*, 154(6):222 (2017). doi:10.3847/1538-3881/aa91d5.

Leistedt, B., Hogg, D. W., Wechsler, R. H., and DeRose, J. Hierarchical modeling and statistical calibration for photometric redshifts. *ArXiv e-prints* (2018).

Leistedt, B., Mortlock, D. J., and Peiris, H. V. Hierarchical Bayesian inference of galaxy redshift distributions from photometric surveys. *MNRAS*, 460:4258–4267 (2016). doi:10.1093/mnras/stw1304.

Lemaître, G. Un Univers homogène de masse constante et de rayon croissant rendant compte de la vitesse radiale des nébuleuses extra-galactiques. *Annales de la Société Scientifique de Bruxelles*, 47:49–59 (1927).

Lemaître, G. Expansion of the universe, a homogeneous universe of constant mass and increasing radius accounting for the radial velocity of extra-galactic nebulae. *Monthly Notices of the Royal Astronomical Society*, 91:483–490 (1931).

Li, Y. Deep Reinforcement Learning: An Overview. *arXiv e-prints*, arXiv:1701.07274 (2017).

Liddle, A. *An introduction to modern cosmology*. John Wiley & Sons (2003).

Lima, M., Cunha, C. E., Oyaizu, H., Frieman, J., Lin, H., and Sheldon, E. S. Estimating the redshift distribution of photometric galaxy samples. *MNRAS*, 390(1):118–130 (2008). doi:10.1111/j.1365-2966.2008.13510.x.

Limber, D. N. The Analysis of Counts of the Extragalactic Nebulae in Terms of a Fluctuating Density Field. *ApJ*, 117:134 (1953). doi:10.1086/145672.

Linder, E. V. Exploring the expansion history of the universe. *Phys. Rev. Lett.*, 90:091301 (2003). doi:10.1103/PhysRevLett.90.091301.

Lindley, D. V. The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337 (2000).

Lipton, Z. C. The Mythos of Model Interpretability. *arXiv e-prints*, arXiv:1606.03490 (2016).

Liu, Q. and Tong, X. T. Accelerating Metropolis-within-Gibbs sampler with localized computations of differential equations. *arXiv e-prints*, arXiv:1906.10541 (2019).

Livio, M. Lost in translation: Mystery of the missing text solved. *Nature*, 479(7372):171 (2011).

Loh, E. D. and Spillar, E. J. Photometric Redshifts of Galaxies. *ApJ*, 303:154 (1986). doi:10.1086/164062.

Loureiro, A., Cuceu, A., Abdalla, F. B., Moraes, B., Whiteway, L., McLeod, M., Balan, S. T., Lahav, O., Benoit-Lévy, A., Manera, M., Rollins, R. P., and Xavier, H. S. Upper Bound of Neutrino Masses from Combined Cosmological Observations and Particle Physics Experiments. *Phys. Rev. Lett.*, 123(8):081301 (2019). doi: 10.1103/PhysRevLett.123.081301.

LSST Science Collaboration, Abell, P. A., Allison, J., Anderson, S. F., Andrew, J. R., Angel, J. R. P., Armus, L., Arnett, D., Asztalos, S. J., Axelrod, T. S., Bailey, S., Ballantyne, D. R., Bankert, J. R., Barkhouse, W. A., Barr, J. D., Barrientos, L. F., Barth, A. J., Bartlett, J. G., Becker, A. C., Becla, J., et al. LSST Science Book, Version 2.0. *arXiv e-prints*, arXiv:0912.0201 (2009).

Lupton, R. H. Sdss image processing i: The deblender. Technical report (2005).

Ly, A., Marsman, M., Verhagen, J., Grasman, R. P., and Wagenmakers, E.-J. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55 (2017). doi:10.1016/j.jmp.2017.05.006.

MacKay, D. J. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472 (1992). doi:10.1162/neco.1992.4.3.448.

Maclaurin, D., Duvenaud, D., and Adams, R. P. Autograd: Effortless gradients in numpy. In *ICML 2015 AutoML Workshop*, volume 238 (2015).

MacQueen, J. Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 281–297. University of California Press, Berkeley, Calif. (1967).

Malz, A. I., Marshall, P. J., DeRose, J., Graham, M. L., Schmidt, S. J., and and, R. W. Approximating photo-z PDFs for large surveys. *The Astronomical Journal*, 156(1):35 (2018). doi:10.3847/1538-3881/aac6b5.

Mandelbaum, R. Weak Lensing for Precision Cosmology. *ARA&A*, 56:393–433 (2018). doi:10.1146/annurev-astro-081817-051928.

Mandelbaum, R., Seljak, U., Hirata, C. M., Bardelli, S., Bolzonella, M., Bongiorno, A., Carollo, M., Contini, T., Cunha, C. E., Garilli, B., Iovino, A., Kampczyk, P., Kneib, J.-P., Knobel, C., Koo, D. C., Lamareille, F., Le Fèvre, O., Leborgne, J.-F., Lilly, S. J., Maier, C., et al. Precision photometric redshift calibration for galaxy–galaxy weak lensing*. *MNRAS*, 386(2):781–806 (2008). ISSN 0035-8711. doi:10.1111/j.1365-2966.2008.12947.x.

Matsumoto, M. and Nishimura, T. Mersenne twister: a 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 8(1):3–30 (1998).

Mayer, J., Khairy, K., and Howard, J. Drawing an elephant with four complex parameters. *American Journal of Physics*, 78(6):648–649 (2010). doi:10.1119/1.3254017.

McLeod, M., Balan, S. T., and Abdalla, F. B. A joint analysis for cosmology and photometric redshift calibration using cross-correlations. *MNRAS*, 466(3):3558–3568 (2017). doi:10.1093/mnras/stw2989.

Mead, A. J., Peacock, J. A., Heymans, C., Joudaki, S., and Heavens, A. F. An accurate halo model for fitting non-linear cosmological power spectra and baryonic feedback models. *MNRAS*, 454(2):1958–1975 (2015). doi:10.1093/mnras/stv2036.

Melchior, P., Moolekamp, F., Jerdee, M., Armstrong, R., Sun, A.-L., Bosch, J., and Lupton, R. SCARLET: Source separation in multi-band images by Constrained Matrix Factorization. *ArXiv e-prints* (2018).

Ménard, B., Scranton, R., Schmidt, S., Morrison, C., Jeong, D., Budavari, T., and Rahman, M. Clustering-based redshift estimation: method and application to data. *ArXiv e-prints* (2013).

Meshcheryakov, A., Glazkova, V., Gerasimov, S., and Mashechkin, I. Measuring the probabilistic photometric redshifts of x-ray quasars based on the quantile regression of ensembles of decision trees. *Astronomy Letters*, 44(12):735–753 (2018). doi:10.1134/S1063773718120058.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.*, 21:1087–1092 (1953). doi:10.1063/1.1699114.

Mink, D. J. and Wyatt, W. F. EMSAO: Radial Velocities from Emission Lines in Spectra. In Shaw, R. A., Payne, H. E., and Hayes, J. J. E., editors, *Astronomical Data Analysis Software and Systems IV*, volume 77 of *Astronomical Society of the Pacific Conference Series*, page 496 (1995).

Murphy, K. P. *Machine learning: a probabilistic perspective*. MIT press (2012).

Myers, A. D., White, M., and Ball, N. M. Incorporating photometric redshift probability density information into real-space clustering measurements. *MNRAS*, 399(4):2279–2287 (2009). ISSN 0035-8711. doi:10.1111/j.1365-2966.2009.15432.x.

Neal, R. M. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media (2012).

Neal, R. M. MCMC using Hamiltonian dynamics. *arXiv e-prints* (2012).

Nesterov, Y. E. A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Dokl. Akad. Nauk SSSR* (1983).

Newman, J. A. Calibrating Redshift Distributions beyond Spectroscopic Limits with Cross-Correlations. *ApJ*, 684:88–101 (2008). doi:10.1086/589982.

Nojiri, S., Odintsov, S. D., and Oikonomou, V. K. Modified gravity theories on a nutshell: Inflation, bounce and late-time evolution. *Phys. Rep.*, 692:1–104 (2017). doi:10.1016/j.physrep.2017.06.001.

Nomoto, K. Accreting white dwarf models for type i supernovae. i. presupernova evolution and triggering mechanisms. *Astrophysical Journal*, 253(2):798–810 (1982).

O'Raifeartaigh, C. and Mitton, S. Interrogating the Legend of Einstein's "Biggest Blunder". *Physics in Perspective*, 20(4):318–341 (2018). doi:10.1007/s00016-018-0228-9.

Ore, O. Pascal and the invention of probability theory. *The American Mathematical Monthly*, 67(5):409–419 (1960).

Osterbrock, D. Active galactic nuclei. *Reports on Progress in Physics*, 54(4):579 (1991).

Palanque-Delabrouille, N., Ruhlmann-Kleider, V., Pascal, S., Rich, J., Guy, J., Bazin, G., Astier, P., Balland , C., Basa, S., Carlberg, R. G., Conley, A., Fouchez, D., Hardin, D., Hook, I. M., Howell, D. A., Pain, R., Perrett, K., Pritchet, C. J., Regnault, N., and Sullivan, M. Photometric redshifts for type Ia supernovae in the supernova legacy survey. *A&A*, 514:A63 (2010). doi:10.1051/0004-6361/200913283.

Palanque-Delabrouille, N., Yèche, C., Baur, J., Magneville, C., Rossi, G., Lesgourgues, J., Borde, A., Burtin, E., LeGoff, J.-M., Rich, J., Viel, M., and Weinberg, D. Neutrino masses and cosmology with Lyman-alpha forest power spectrum. *J. Cosmology Astropart. Phys.*, 2015(11):011 (2015). doi:10.1088/1475-7516/2015/11/011.

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. Automatic differentiation in pytorch (2017).

Pauli, W. Letter to l. In *Meitner and her colleagues (letter open to the participants of the conference in Tübingen)* (1930).

Peacock, J. A. and Smith, R. E. Halo occupation numbers and galaxy bias. *MNRAS*, 318(4):1144–1156 (2000). doi:10.1046/j.1365-8711.2000.03779.x.

Peebles, P. J. and Ratra, B. The cosmological constant and dark energy. *Reviews of Modern Physics*, 75(2):559–606 (2003). doi:10.1103/RevModPhys.75.559.

Peebles, P. J. E. The Galaxy and Mass N-Point Correlation Functions: a Blast from the Past. In Martínez, V. J., Trimble, V., and Pons-Bordería, M. J., editors, *Historical Development of Modern Cosmology*, volume 252 of *Astronomical Society of the Pacific Conference Series*, page 201 (2001).

Perlmutter, S., Aldering, G., Goldhaber, G., Knop, R. A., Nugent, P., Castro, P. G., Deustua, S., Fabbro, S., Goobar, A., Groom, D. E., Hook, I. M., Kim, A. G., Kim, M. Y., Lee, J. C., Nunes, N. J., Pain, R., Pennypacker, C. R., Quimby, R., Lidman, C., Ellis, R. S., et al. Measurements of Ω and Λ from 42 High-Redshift Supernovae. *ApJ*, 517(2):565–586 (1999). doi:10.1086/307221.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook (2014).

Petri, A., May, M., and Haiman, Z. Cosmology with photometric weak lensing surveys: Constraints with redshift tomography of convergence peaks and moments. *Phys. Rev. D*, 94(6):063534 (2016). doi:10.1103/PhysRevD.94.063534.

Phillips, M. M. The absolute magnitudes of Type IA supernovae. *ApJ*, 413:L105–L108 (1993). doi:10.1086/186970.

Phillips, M. M., Lira, P., Suntzeff, N. B., Schommer, R. A., Hamuy, M., and Maza, J. The Reddening-Free Decline Rate Versus Luminosity Relationship for Type IA Supernovae. *AJ*, 118(4):1766–1776 (1999). doi:10.1086/301032.

Planck Collaboration, Ade, P. A. R., Aghanim, N., Arnaud, M., Ashdown, M., Aumont, J., Baccigalupi, C., Banday, A. J., Barreiro, R. B., Bartlett, J. G., and et al. Planck 2015 results. XIII. Cosmological parameters. *A&A*, 594:A13 (2016). doi:10.1051/0004-6361/201525830.

Planck Collaboration, Aghanim, N., Akrami, Y., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., et al. Planck 2018 results. VI. Cosmological parameters. *arXiv e-prints*, arXiv:1807.06209 (2018a).

Planck Collaboration, Akrami, Y., Arroja, F., Ashdown, M., Aumont, J., Baccigalupi, C., Ballardini, M., Banday, A. J., Barreiro, R. B., Bartolo, N., Basak, S., Battye, R., Benabed, K., Bernard, J. P., Bersanelli, M., Bielewicz, P., Bock, J. J., Bond, J. R., Borrill, J., Bouchet, F. R., et al. Planck 2018 results. I. Overview and the cosmological legacy of Planck. *arXiv e-prints*, arXiv:1807.06205 (2018b).

Pound, R. V. and Rebka, G. A. Gravitational red-shift in nuclear resonance. *Phys. Rev. Lett.*, 3:439–441 (1959). doi:10.1103/PhysRevLett.3.439.

Prechelt, L. *Early stopping-but when?* Springer (1998). doi:10.1007/3-540-49430-8_3.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press (2007).

Qian, X. and Vogel, P. Neutrino mass hierarchy. *Progress in Particle and Nuclear Physics*, 83:1–30 (2015). doi:10.1016/j.ppnp.2015.05.002.

Quadri, R. F. and Williams, R. J. Quantifying photometric redshift errors in the absence of spectroscopic redshifts. *The Astrophysical Journal*, 725(1):794 (2010).

Racca, G. D., Laureijs, R., Stagnaro, L., Salvignol, J.-C., Lorenzo Alvarez, J., Saavedra Criado, G., Gaspar Venancio, L., Short, A., Strada, P., Bönke, T., Colombo, C.,

Calvi, A., Maiorano, E., Piersanti, O., Prezelus, S., Rosato, P., Pinel, J., Rozemeijer, H., Lesna, V., Musi, P., et al. The Euclid mission design. In *Space Telescopes and Instrumentation 2016: Optical, Infrared, and Millimeter Wave*, volume 9904 of *Proc. SPIE*, page 99040O (2016). doi:10.1117/12.2230762.

Rahman, M., Mendez, A. J., Ménard, B., Scranton, R., Schmidt, S. J., Morrison, C. B., and Budavári, T. Exploring the sdss photometric galaxies with clustering redshifts. *Monthly Notices of the Royal Astronomical Society*, 460(1):163–174 (2016).

Ramond, P. Neutrinos: a glimpse beyond the standard model. *Nuclear Physics B Proceedings Supplements*, 77:3–9 (1999).

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. *arXiv preprint arXiv:1401.0118* (2013).

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press (2005). ISBN 026218253X.

Rau, M. M., Wilson, S., and Mandelbaum, R. Estimating redshift distributions using Hierarchical Logistic Gaussian processes. *arXiv e-prints*, arXiv:1904.09988 (2019).

Ravanbakhsh, S., Oliva, J., Fromenteau, S., Price, L. C., Ho, S., Schneider, J., and Poczos, B. Estimating Cosmological Parameters from the Dark Matter Distribution. *arXiv e-prints*, arXiv:1711.02033 (2017).

Reiman, D. M. and Göhre, B. E. Deblending galaxy superpositions with branched generative adversarial networks. *MNRAS*, 485(2):2617–2627 (2019). doi:10.1093/mnras/stz575.

Rhodes, J., Nichol, R. C., Aubourg, É., Bean, R., Boutigny, D., Bremer, M. N., Capak, P., Cardone, V., Carry, B., Conselice, C. J., Connolly, A. J., Cuillandre, J.-C., Hatch, N. A., Helou, G., Hemmati, S., Hildebrandt, H., Hložek, R., Jones, L., Kahn, S., Kiessling, A., et al. Scientific Synergy between LSST and Euclid. *ApJS*, 233:21 (2017). doi:10.3847/1538-4365/aa96b0.

Riess, A. G., Filippenko, A. V., Challis, P., Clocchiatti, A., Diercks, A., Garnavich, P. M., Gilliland, R. L., Hogan, C. J., Jha, S., Kirshner, R. P., Leibundgut, B., Phillips, M. M., Reiss, D., Schmidt, B. P., Schommer, R. A., Smith, R. C., Spyromilio, J., Stubbs, C., Suntzeff, N. B., and Tonry, J. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *AJ*, 116:1009–1038 (1998). doi:10.1086/300499.

Riess, A. G., Macri, L. M., Hoffmann, S. L., Scolnic, D., Casertano, S., Filippenko, A. V., Tucker, B. E., Reid, M. J., Jones, D. O., Silverman, J. M., Chornock, R., Challis, P., Yuan, W., Brown, P. J., and Foley, R. J. A 2.4% Determination of the Local Value of the Hubble Constant. *ApJ*, 826(1):56 (2016). doi:10.3847/0004-637X/826/1/56.

Rivera, J. D., Moraes, B., Merson, A. I., Jouvel, S., Abdalla, F. B., and Abdalla, M. C. B. Degradation analysis in the estimation of photometric redshifts from non-representative training sets. *MNRAS*, 477:4330–4347 (2018). doi:10.1093/mnras/sty880.

Robertson, H. P. Kinematics and World-Structure. *ApJ*, 82:284 (1935). doi:10.1086/143681.

Rodríguez-Muñoz, L., Rodighiero, G., Mancini, C., Pérez-González, P., Rawle, T., Egami, E., Mercurio, A., Rosati, P., Puglisi, A., Franceschini, A., et al. Quantifying the suppression of the (un)-obscured star formation in galaxy cluster cores at 0.2 z 0.9. *Monthly Notices of the Royal Astronomical Society*, 485(1):586–619 (2019). doi:10.1093/mnras/sty3335.

Ross, A. J., Samushia, L., Howlett, C., Percival, W. J., Burden, A., and Manera, M. The clustering of the SDSS DR7 main Galaxy sample - I. A 4 per cent distance measure at z = 0.15. *MNRAS*, 449:835–847 (2015). doi:10.1093/mnras/stv154.

Rudd, D. H., Zentner, A. R., and Kravtsov, A. V. Effects of baryons and dissipation on the matter power spectrum. *The Astrophysical Journal*, 672(1):19 (2008).

Ruder, S. An overview of gradient descent optimization algorithms. *arXiv e-prints*, arXiv:1609.04747 (2016).

Rudin, C. Please Stop Explaining Black Box Models for High Stakes Decisions. *arXiv e-prints*, arXiv:1811.10154 (2018).

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. Learning representations by back-propagating errors. *Nature*, 323:533–536 (1986). doi:10.1038/323533a0.

Runnalls, A. R. Kullback-leibler approach to gaussian mixture reduction. *IEEE Transactions on Aerospace and Electronic Systems*, 43(3):989–999 (2007). ISSN 0018-9251. doi:10.1109/TAES.2007.4383588.

Sadeh, I., Abdalla, F. B., and Lahav, O. ANNz2: Photometric Redshift and Probability Distribution Function Estimation using Machine Learning. *PASP*, 128(10):104502 (2016). doi:10.1088/1538-3873/128/968/104502.

Salvatier, J., Wiecki, T. V., and Fonnesbeck, C. Probabilistic programming in python using PyMC3. *PeerJ Computer Science*, 2:e55 (2016). doi:10.7717/peerj-cs.55.

Salvato, M., Ilbert, O., and Hoyle, B. The many flavours of photometric redshifts. *Nature Astronomy*, 3(3):212 (2019).

Samuroff, S., Troxel, M. A., Bridle, S. L., Zuntz, J., MacCrann, N., Krause, E., Eifler, T., and Kirk, D. Simultaneous constraints on cosmology and photometric redshift bias from weak lensing and galaxy clustering. *MNRAS*, 465:L20–L24 (2017). doi:10.1093/mnrasl/slw201.

Sánchez, C. and Bernstein, G. M. Redshift inference from the combination of galaxy colours and clustering in a hierarchical Bayesian model. *MNRAS*, 483(2):2801–2813 (2019). doi:10.1093/mnras/sty3222.

Sánchez, C., Carrasco Kind, M., Lin, H., Miquel, R., Abdalla, F. B., Amara, A., Banerji, M., Bonnett, C., Brunner, R., Capozzi, D., et al. Photometric redshift

analysis in the dark energy survey science verification data. *Monthly Notices of the Royal Astronomical Society*, 445(2):1482–1506 (2014).

Schieferdecker, D. and Huber, M. F. Gaussian mixture reduction via clustering. In *2009 12th International Conference on Information Fusion*, pages 1536–1543. IEEE (2009).

Schmidt, S. J., Malz, A. I., Soo, J. Y. H., Almosallam, I. A., Brescia, M., Cavuoti, S., Cohen-Tanugi, J., Connolly, A. J., DeRose, J., Freeman, P. E., Graham, M. L., Iyer, K. G., Jarvis, M. J., Kalmbach, J. B., Kovacs, E., Lee, A. B., Longo, G., Morrison, C. B., Newman, J. A., Nourbakhsh, E., et al. Evaluation of probabilistic photometric redshift estimation approaches for LSST. *arXiv e-prints*, arXiv:2001.03621 (2020).

Schmidt, S. J., Ménard, B., Scranton, R., Morrison, C., and McBride, C. K. Recovering redshift distributions with cross-correlations: pushing the boundaries. *MNRAS*, 431:3307–3318 (2013). doi:10.1093/mnras/stt410.

Schmidt, S. J. and Thorman, P. Improved photometric redshifts via enhanced estimates of system response, galaxy templates and magnitude priors. *MNRAS*, 431:2766–2777 (2013). doi:10.1093/mnras/stt373.

Schneider, A., Teyssier, R., Stadel, J., Chisari, N. E., Le Brun, A. M. C., Amara, A., and Refregier, A. Quantifying baryon effects on the matter power spectrum and the weak lensing shear correlation. *J. Cosmology Astropart. Phys.*, 2019(3):020 (2019). doi:10.1088/1475-7516/2019/03/020.

Scottez, V., Benoit-Lévy, A., Coupon, J., Ilbert, O., and Mellier, Y. Testing the accuracy of clustering redshifts with simulations. *MNRAS*, 474(3):3921–3930 (2018). doi:10.1093/mnras/stx3056.

Seljak, U. Analytic model for galaxy and dark matter clustering. *MNRAS*, 318:203–213 (2000). doi:10.1046/j.1365-8711.2000.03715.x.

Sellentin, E. and Starck, J.-L. Debiasing inference with approximate covariance matrices and other unidentified biases. *arXiv e-prints*, arXiv:1902.00709 (2019).

Semboloni, E., Tereno, I., van Waerbeke, L., and Heymans, C. Sources of contamination to weak lensing tomography: redshift-dependent shear measurement bias. *MNRAS*, 397:608–622 (2009). doi:10.1111/j.1365-2966.2009.14926.x.

Shapiro, I. I. Fourth test of general relativity. *Physical Review Letters*, 13(26):789 (1964).

Shrikumar, A., Greenside, P., and Kundaje, A. Learning Important Features Through Propagating Activation Differences. *arXiv e-prints*, arXiv:1704.02685 (2017).

Shwartz-Ziv, R. and Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv e-prints*, arXiv:1703.00810 (2017).

Sivia, D. and Skilling, J. *Data analysis: a Bayesian tutorial*. OUP Oxford (2006).

Skilling, J. Nested sampling for general bayesian computation. *Bayesian Anal.*, 1(4):833–859 (2006). doi:10.1214/06-BA127.

Slipher, V. M. Radial velocity observations of spiral nebulae. *The Observatory*, 40:304–306 (1917).

Smith, R. E., Peacock, J. A., Jenkins, A., White, S. D. M., Frenk, C. S., Pearce, F. R., Thomas, P. A., Efstathiou, G., and Couchman, H. M. P. Stable clustering, the halo model and non-linear cosmological power spectra. *MNRAS*, 341(4):1311–1332 (2003). doi:10.1046/j.1365-8711.2003.06503.x.

Solà, J. and Štefančić, H. Effective equation of state for dark energy: Mimicking quintessence and phantom energy through a variable λ. *Physics letters B*, 624(3-4):147–157 (2005).

Sołtan, A. M. Evolution of the galaxy correlation function at redshifts $0.2 < z < 3$. In van de Weygaert, R., Shandarin, S., Saar, E., and Einasto, J., editors, *The Zeldovich Universe: Genesis and Growth of the Cosmic Web*, volume 308 of *IAU Symposium*, pages 291–292 (2016). doi:10.1017/S1743921316010000.

Soo, J. Y. H., Moraes, B., Joachimi, B., Hartley, W., Lahav, O., Charbonnier, A., Makler, M., Pereira, M. E. S., Comparat, J., Erben, T., Leauthaud, A., Shan, H., and Van Waerbeke, L. Morpho-z: improving photometric redshifts with galaxy morphology. *MNRAS*, 475:3613–3632 (2018). doi:10.1093/mnras/stx3201.

Sotiriou, T. P. and Faraoni, V. f(R) theories of gravity. *Reviews of Modern Physics*, 82(1):451–497 (2010). doi:10.1103/RevModPhys.82.451.

Speagle, J. S. and Eisenstein, D. J. Deriving photometric redshifts using fuzzy archetypes and self-organizing maps - I. Methodology. *MNRAS*, 469:1186–1204 (2017). doi:10.1093/mnras/stw1485.

Sprott, D. A. *Statistical inference in science*. Springer Science & Business Media (2008).

Stabenau, H. F., Connolly, A., and Jain, B. Photometric redshifts with surface brightness priors. *Monthly Notices of the Royal Astronomical Society*, 387(3):1215–1226 (2008).

Stoughton, C., Lupton, R. H., Bernardi, M., Blanton, M. R., Burles, S., Castander, F. J., Connolly, A. J., Eisenstein, D. J., Frieman, J. A., Hennessy, G. S., Hindsley, R. B., Ivezić, Ž., Kent, S., Kunszt, P. Z., Lee, B. C., Meiksin, A., Munn, J. A., Newberg, H. J., Nichol, R. C., Nicinski, T., et al. Sloan Digital Sky Survey: Early Data Release. *AJ*, 123:485–548 (2002). doi:10.1086/324741.

Sun, L., Zhan, H., and Tao, C. Reconstructing the galaxy redshift distribution from angular cross power spectra. *ArXiv e-prints* (2015).

Syring, N. and Martin, R. Calibrating general posterior credible regions. *Biometrika*, 106(2):479–486 (2018). doi:10.1093/biomet/asy054.

Tanaka, M. Photometric Redshift with Bayesian Priors on Physical Properties of Galaxies. *ApJ*, 801:20 (2015). doi:10.1088/0004-637X/801/1/20.

Taylor, J. H. and Weisberg, J. M. A new test of general relativity - Gravitational radiation and the binary pulsar PSR 1913+16. *ApJ*, 253:908–920 (1982). doi:10.1086/159690.

Taylor, P. L., Kitching, T. D., Alsing, J., Wandelt, B. D., Feeney, S. M., and McEwen, J. D. Cosmic shear: Inference from forward models. *Phys. Rev. D*, 100(2):023519 (2019). doi:10.1103/PhysRevD.100.023519.

The LSST Dark Energy Science Collaboration, Mandelbaum, R., Eifler, T., Hložek, R., Collett, T., Gawiser, E., Scolnic, D., Alonso, D., Awan, H., Biswas, R., Blazek, J., Burchat, P., Chisari, N. E., Dell'Antonio, I., Digel, S., Frieman, J., Goldstein, D. A., Hook, I., Ivezić, Ž., Kahn, S. M., et al. The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document. *arXiv e-prints*, arXiv:1809.01669 (2018).

Theuns, T. Physical cosmology - lecture notes (2016).

Tonry, J. and Davis, M. A survey of galaxy redshifts. i-data reduction techniques. *The Astronomical Journal*, 84:1511–1525 (1979).

Trivedi, P. K., Zimmer, D. M., et al. Copula modeling: an introduction for practitioners. *Foundations and Trends® in Econometrics*, 1(1):1–111 (2007).

Trotta, R. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49:71–104 (2008). doi:10.1080/00107510802066753.

Troxel, M. A. and Ishak, M. The intrinsic alignment of galaxies and its impact on weak gravitational lensing in an era of precision cosmology. *Phys. Rep.*, 558:1–59 (2015). doi:10.1016/j.physrep.2014.11.001.

Troxel, M. A., MacCrann, N., Zuntz, J., Eifler, T. F., Krause, E., Dodelson, S., Gruen, D., Blazek, J., Friedrich, O., Samuroff, S., Prat, J., Secco, L. F., Davis, C., Ferté, A., DeRose, J., Alarcon, A., Amara, A., Baxter, E., Becker, M. R., Bernstein, G. M., et al. Dark Energy Survey Year 1 Results: Cosmological Constraints from Cosmic Shear. *ArXiv e-prints* (2017).

Turner, B. M. and Zandt, T. V. A tutorial on approximate bayesian computation. *Journal of Mathematical Psychology*, 56(2):69 – 85 (2012). ISSN 0022-2496. doi: https://doi.org/10.1016/j.jmp.2012.02.005.

Wadadekar, Y. Estimating Photometric Redshifts Using Support Vector Machines. *PASP*, 117(827):79–85 (2005). doi:10.1086/427710.

Walker, A. G. On Milne's Theory of World-Structure. *Proceedings of the London Mathematical Society, (Series 2) volume 42, p. 90-127*, 42:90–127 (1937). doi:10.1112/plms/s2-42.1.90.

Way, M. J. and Srivastava, A. N. Novel Methods for Predicting Photometric Redshifts from Broadband Photometry Using Virtual Sensors. *ApJ*, 647:102–115 (2006). doi:10.1086/505293.

Weir, N., Fayyad, U. M., Djorgovski, S. G., and Roden, J. The SKICAT System for Processing and Analyzing Digital Imaging Sky Surveys. *PASP*, 107:1243 (1995). doi:10.1086/133683.

Weisberg, J. M., Nice, D. J., and Taylor, J. H. Timing Measurements of the Relativistic Binary Pulsar PSR B1913+16. *ApJ*, 722(2):1030–1034 (2010). doi:10.1088/0004-637X/722/2/1030.

West, M. Approximating posterior distributions by mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):409–422 (1993). doi:10.1111/j.2517-6161.1993.tb01911.x.

Williams, J. L. and Maybeck, P. S. Cost-function-based hypothesis control techniques for multiple hypothesis tracking. *Mathematical and Computer Modelling*, 43(9):976 – 989 (2006). ISSN 0895-7177. doi:https://doi.org/10.1016/j.mcm.2005.05.022. Optimization and Control for Military Applications.

Wittman, D., Bhaskar, R., and Tobin, R. Overconfidence in photometric redshift estimation. *MNRAS*, 457:4005–4011 (2016). doi:10.1093/mnras/stw261.

Wittman, D., Riechers, P., and Margoniner, V. E. Photometric Redshifts and Photometry Errors. *ApJ*, 671(2):L109–L112 (2007). doi:10.1086/525020.

Yasuda, N., Fukugita, M., Narayanan, V. K., Lupton, R. H., Strateva, I., Strauss, M. A., Ivezić, Ž., Kim, R. S. J., Hogg, D. W., Weinberg, D. H., Shimasaku, K., Loveday, J., Annis, J., Bahcall, N. A., Blanton, M., Brinkmann, J., Brunner, R. J., Connolly, A. J., Csabai, I., Doi, M., et al. Galaxy Number Counts from the Sloan Digital Sky Survey Commissioning Data. *AJ*, 122:1104–1124 (2001). doi:10.1086/322093.

Permission to Reproduce Figures

Figures 2.4, 2.5 and 5.1 have been reproduced from *Monthly Notices of the Royal Astronomical Society*. Explicit permission is not required for their reproduction for the purpose of academic research[1].

Figure 2.3 has been reproduced from *Astronomy and Astrophysics*. Permission to reproduce this figure was requested and granted, as shown on the following page.

[1] `https://academic.oup.com/mnras/pages/rights_and_new_business_development`

Astronomy and Astrophysics
Editor in Chief: T. Forveille

T. Forveille
Astronomy & Astrophysics
Observatoire de Paris
61, avenue de l'Observatoire
75014 Paris, France

Tel.: 33 0(1) 43 29 05 41
Fax: 33 0(1) 43 29 05 57
e-mail: aanda.paris@obspm.fr
Web: http://www.aanda.org

merging
Annales d'Astrophysique
Arkiv for Astronomi
Bulletin of the Astronomical Institutes of the Netherlands
Bulletin Astronomique
Journal des Observateurs
Zeitschrift fur Astrophysik
Bulletin of the Astronomical Institutes of Czechoslovakia

Paris, September 23, 2019

Reprint Permission

Material:
Fig. 1 in Planck Collaboration 2016, A&A, 594, A13

To be used in:
PhD thesis entitled 'Photometric Redshifts for Future Cosmological Galaxy Surveys' at Imperial College London

Permission granted to:
Daniel Jones
d.jones15@imperial.ac.uk

I hold copyright on the material referred to above, and hereby grant permission for its use as requested herewith. Credit should be given as follows:
Credit: Author, A&A, vol, page, year, reproduced with permission © ESO.

Thierry Forveille
A&A Editor-in-Chief

Sponsored by Argentina, Armenia, Austria, Belgium, Bulgaria, Chile, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Slovak Republic, Spain, Sweden, and Switzerland.
Produced and distributed by EDP Sciences for ESO.