Dark Energy Science from 100 Million Galaxies: AI-Driven Analysis and Data-Intensive Techniques for Cosmological Discovery

Andresa Rodrigues de Campos

Department of Physics
Carnegie Mellon University

Thesis Committee:
Scott Dodelson (Chair)
Roy Briere
Rachel Mandelbaum
Chad Schafer

December, 2023

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

Abstract

Observational cosmology is a rapidly evolving field. Thanks to technological advancements, the advent of big data, machine learning, and international collaborations, there have been significant advances in cosmology in recent years, which have greatly enhanced our understanding of the universe. Observational cosmology aims to thoroughly test theoretical predictions about the expansion history of the universe and the evolution of cosmic structure over time. This is achieved through cosmological surveys associated with a variety of observables. Measurements derived from sources such as the cosmic microwave background (CMB), exemplified by the *Planck* satellite's detailed mapping of the CMB's temperature fluctuations, and the distance-redshift relationship using Type Ia supernovae, as observed in projects like the Supernova Legacy Survey, provide essential data. Baryonic acoustic oscillations (BAO) observed in the clustering of galaxies, such as those charted by the Sloan Digital Sky Survey (SDSS), along with the observed growth of cosmic structure through galaxy clustering and gravitational lensing phenomena, as investigated by surveys like the Dark Energy Survey (DES), the Kilo-Degree Survey (KiDS) and the Hyper Suprime-Cam (HSC), all contribute to a coherent picture.

The collective evidence from these surveys indicates that deviations from the predictions of the ΛCDM (Lambda Cold Dark Matter) standard cosmological model are minor, typically within a few percent. However, the next phase in this research program is to achieve even greater precision and accuracy in our measurements to robustly challenge the ΛCDM model with empirical data. Current and upcoming experiments, such as the Euclid mission, the Nancy Grace Roman Space Telescope and the Vera C. Rubin Observatory's Legacy Survey of Space and Time (LSST), have been meticulously designed to reduce statistical uncertainties in cosmological measurements, aiming to surpass the current state of the art. Nonetheless, assuming that these surveys successfully gather data, it is anticipated that the primary challenges in our quest for deeper cosmological insights will arise from systematic uncertainties. Thus, the future challenges we face are not solely about improving statistical precision but also involve identifying and mitigating sources of systematics that could influence the accuracy and integrity of our cosmological findings. This thesis explores several crucial facets pertaining to systematic uncertainties in cosmological inquiries. Moreover, considering that the concordance of predictions across different surveys is essential for validating the cosmological model, this thesis also encompasses a critical examination of inter-survey consistency.

The first major emphasis of this study is the mitigation of systematics associated with photometric redshift estimation. An accurate characterization of the redshift distribution, $n(z)$, for the observed sample is crucial for cosmological analyses, particularly in the context of weak lensing shear studies. To this end, I have improved the Self-Organizing Map (SOM) method for photometric redshift estimation, which I refer to as SOMPZ. This approach, which leverages unsupervised machine learning, was initially implemented for the DES Year 3 (DES Y3). I have further enhanced it for the upcoming DES Y6 data set. The analyses in this thesis show substantial improvements by substituting the Y3 SOM algorithm with an optimized version that better addresses the intricacies of redshift estimation. Moreover, the integration of g-band flux data has markedly enhanced redshift precision, achieving a reduction in the overlap between redshift bins by as much as 66%. These advancements are key in refining weak lensing redshift characterization, setting a higher standard not just for DES Y6, but also for future stage IV surveys like the Rubin Observatory.

The second pivotal subject of this thesis is an empirical approach to model selection, with a focus on explicitly balancing parameter bias against model complexity. This approach utilizes synthetic data to calibrate the relationship between bias and the χ^2 difference between models. It enables the interpretation of χ^2 values obtained from real data, even when catalogs are blinded, facilitating informed decisions regarding model selection. This method is applied to tackle the challenge of intrinsic alignments, a significant systematic uncertainty in weak lensing studies that substantially contributes to the error budget in modern lensing surveys. Specifically, I compare two commonly used models, nonlinear alignment (NLA) and tidal alignment & tidal torque (TATT), against bias in the $\Omega_{\rm m} - S_8$ plane, with a particular focus on the DES Y3. In this case, there is a roughly a 30% chance that were NLA to be the fiducial model, the results would be biased (in the $\Omega_{\rm m} - S_8$ plane) by more than 0.3σ .

Lastly, the third focus of this thesis involves the application of several tension estimators to assess the DES large-scale structure measurement and *Planck* cosmic microwave background data. These tension metrics are evaluated for their responsiveness to artificially introduced tension between the two data sets using synthetic DES data. Given the importance of tensions, which represent discrepancies in cosmological parameter measurements across different experiments, identifying them is critical. Statistical significant tensions may hint at novel physics beyond the standard cosmological model, or unaccounted systematics. These tension metrics are then applied to compare *Planck* and actual DES Y1 data. The parameter differences, Eigentension, and Suspiciousness metrics yield consistent results on both simulated and real data, while the Bayes ratio stands out due to its dependence on the prior volume. Using these metrics, we calculate the tension between DES Y1 3×2pt and *Planck* revealing that the surveys are in approximately 2.3σ tension under the ΛCDM paradigm. This suite of metrics provided a robust tool set for testing tensions in the DES Y3 data, where we found approximately 0.7σ tension to *Planck* 2018 under the ΛCDM paradigm.

In summary, the projects that compose this thesis are dedicated to the development and enhancement of statistical and machine learning methodologies for the analysis of extensive data sets in large-scale structure surveys.

Acknowledgments

First and foremost, I extend my deepest gratitude to my advisor, Professor Scott Dodelson, whose support, guidance, and mentorship have been foundational to my research journey. My first encounter with cosmology was through Scott's book, and it has been an honor to later become his Ph.D. student and learn from him directly over these past years. His keen insights and profound wisdom have been instrumental in shaping this thesis and have sharpened my abilities as a researcher. I particularly admire Scott's skill in distilling complex problems to their essence, paving the way for clear thinking and effective solutions. His mentorship has been invaluable in broadening my research network and guiding my path within the Dark Energy Survey collaboration, while also providing me ample freedom to pursue my research interests. Most importantly, I am profoundly grateful to Scott for his patience and understanding whenever I faced challenges, and for the flexibility he has afforded me in completing this Ph.D.

My gratitude also goes out to my thesis committee members, Professor Roy Briere, Professor Rachel Mandelbaum, and Professor Chad Schafer. Their valuable and insightful feedback has significantly contributed to my growth as a researcher. I thank Rachel for her additional mentorship during our collaborative project, the enriching opportunity to co-organize the McWilliams Software Seminar Series, and for teaching me so much about weak lensing in our weekly group meetings. I also owe many thanks to Professors Roy Briere and Chad Schafer for their constructive suggestions that have enhanced my research projects. Lastly, I extend my sincere thanks to each member of the committee for the time and effort taken to read through my thesis. Their advice and encouragement during my annual reviews, coupled with their careful reading of my work, have been crucial to the successful completion of my thesis.

I wish to express my profound gratitude to my co-authors and colleagues from the Dark Energy Survey, for their significant contributions to my research and growth as a scientist. I would like to thank Simon Samuroff for so much support throughout my entire Ph.D., including answering so many of my questions about weak lensing and intrinsic alignments, and for being always willing to help me. Having his support made this journey much more enjoyable. A huge thanks to Lucas Secco, who helped me so much to get started with everything I needed to start working in the DES (CosmoSIS, chains, plotting, etc.), and worked with me on my first DES project, the analysis assuming w0wa, which was probably responsible for opening many doors for me. I also owe my gratitude to Alexandra Amon, who has supported me in so many ways. Alex taught me so much about weak lensing and photometric redshifts, and has provided so much guidance and feedback, by carefully reading my work, providing suggestions, and making time for many one-to-one calls to clarify my questions. Many thanks to Judit Prat, Alex Alarcon, and Carles Sanchez, who made me feel welcomed in DES since the very beginning. After meeting them at the Michigan workshop, it felt like we had been working together for years. Actually working with them, in the following years, has been an absolute joy. Thank you Giulia Giannini and Marco Gatti, for being such great collaborators, always willing to help and answering my questions, and being so fun to hang out with. I am particularly in debt with Giulia for always having my back whenever I was busy. A heartfelt thanks to Boyan Yin, for working with me on redshifts over the pandemic years and great distances. She is an amazing collaborator, and I am lucky to have been able to count on

her. Finally, my profound gratitude to all the remarkable individuals in the Dark Energy Survey, including Justin Myles, Anna Porredon, Jessie Muir, Noah Weaverdyck, Agnès Ferté, Dane Cross, David Cid, Masaya Yamamoto, Pablo Lemos, Marco Raveri, Dragan Huterer, Michael Troxel, Matthew Becker, Gary Bernstein, Chihway Chang, Jonathan Blazek, Daniel Gruen, Mike Jarvis, Bhuv Jain, and many others. Their exceptional expertise and collaborative spirit have not only advanced our projects but have also been a constant source of inspiration and learning for me.

I am immensely grateful for the wonderful colleagues and friends I have found in the Physics Department at CMU. It has been a privilege to embark on this Ph.D. journey with such an incredible cohort, including Abhirami Harilal, I-Hsuan Kao, Ryan Muzzio, Kuldeep Sharma, Beka Modrekiladze, Nianyi Chen and Tianqing Zhang. Their camaraderie during our intense first and second years, our shared struggles and triumphs over challenging assignments, and the countless hours we spent together as TAs have been a crucial part of my academic journey. I also extend my heartfelt thanks to those with whom I shared those much-needed coffee breaks at 2 o'clock, including Larisa Thorne, Alejandro Sanchez, Malavika Varma, Markus Deserno, Teo, and Manfred Paulini. Our casual conversations and the laughter we shared during those breaks were the refreshments I needed to reenergize and tackle the challenges of the day. Special appreciation goes to my cosmology group at CMU, including Danielle Leonard, Husni Almoubayyed, Peikai Li, Junzhe Zhou, Andy Park, Federico Berlfein, Xiangchong Li, and Alex Malz. Their passion for cosmology and willingness to engage in thoughtful discussions, offer feedback on plots, and celebrate each other's achievements have been incredibly motivating. Your commitment to our shared field of study has not only fostered a vibrant academic environment but has also been a constant source of inspiration and support throughout my Ph.D. journey

This thesis would not have been possible without the love and support of my family. First and foremost I want to thank the love of my life and husband, Manuel Olguín Muñoz. His optimistic and cheerful attitude has been a beacon of light, especially in moments of doubt and discouragement. I could not have done this without his companionship and love. Despite the ocean that often laid between us, he made every effort to ensure that I never felt alone on this journey. I am profoundly and eternally grateful for his unconditional support, for always taking care of me, and for making every moment we share together extraordinary. Eu também quero agradecer aos meus pais, Marinete Rodrigues Macario e Marcos Roberto de Campos, por sempre acreditarem em mim. Eles sempre confiaram nas minhas escolhas, atém mesmo sem enteder-las algumas vezes, e sempre fizeram o seu melhor para me apoiar, para que eu pudesse focar em estudar sem outras preocupações – chegar até aqui só foi possível graças a vocês, obrigada. My deepest gratitude goes also to my brother, Marcos Fernando Rodrigues de Campos, for always supporting me. I am forever grateful to him for taking care of our parents. Knowing that I can always count on him makes everything easier. A heartfelt thank you goes also to my parents-in-law, Valeria Muñoz Mendel and Gabriel Olguín Parada, for welcoming me as part of their family, and for their endless encouragement and support. I want to also thank my sister-in-law, Paola, for all the great anime recommendations and for being an amazingly supportive friend.

I remain forever grateful to my dear friend, Natalia Tenorio Maia, for all her support when I first set foot on this new path. Her friendship and encouragement during those initial days were invaluable and I am lucky to have such a kind friend. I also extend my heartfelt thanks to my mentor, Rogerio Rosenfeld, and friends from the Instituto de Física Teórica (IFT), Vinicius Terra, Leonidas Prado (and Camila Coelho), Jogean Ferreira, Victor César, Caroline Costa, Otávio Alves, Felipe Oliveira and Antonino Troja. Their support and camaraderie in the period leading up to my Ph.D. were instrumental in paving the way for this academic pursuit. The solidarity and insights shared during our time at IFT have been fundamental to my growth and confidence as I embarked on this journey. A special note of gratitude is reserved for Daniel Lombelo Teixeira. Daniel's support at the outset of this journey was fundamental. His encouragement and motivation enabled me to embark on this path. For this, and for the countless ways in which he was a pillar of support, I am forever grateful.

My time at CMU was made infinitely brighter by the friends I made through Tartan Salsa, including Manuel Olguín Muñoz, Jennifer Bone, Alyson Barra, Efraín Guirola, Kuai-Kuai Jin, Eric Huang, Moataz Abdulhafez, Nora Abuelil, Carmen Fisac, Russel Brown, and many others. Their companionship transformed Pittsburgh into a place that felt like home. Dancing away the stress of my Ph.D. with them was the perfect escape from the rigors of academic life. Their friendship has infused my days with a vibrancy and energy that I will always cherish. A heartfelt thank you to Kuai-Kuai, who, during a year of isolation, was not just a friend but a lifeline. Her kindness in driving me to get my COVID vaccine and the countless video calls helped maintain my sanity in those trying times. I am also profoundly grateful to Jenn and Alyson for their invaluable help in organizing my elopement ceremony. Their efforts in making that day special and memorable have left an indelible mark on my heart.

I would like to extend a special acknowledgment to my friends from the Brazilian cosmology community, including Vivian Miranda, Lucas Secco, Pedro Bernadinelli (and Val Souza), Nickolas Kokron, Gabriela Sato-Polito, and Karolina García. Their camaraderie and our shared passion for the wonders of cosmology have been a source of inspiration and happiness in my journey. The discussions we have had, the ideas we have shared, and the mutual encouragement in our pursuit of understanding the cosmos have not only propelled my professional growth but have also brought immense joy and a sense of community to my life. Their friendship and dedication to our field are treasures I hold dear.

Last but certainly not least, my deepest gratitude goes to my beloved cats, Mochi and Misú. I thank them for being more than just pets; for being my unwavering companions through the highs and lows of this journey. The comfort, joy, and unconditional love they provide are beyond words. To Misú, with her playful silliness, and to Mochi, with her calming presence, they both have been sources of solace and happiness in my life. They are not just my pets, but my cherished babies, and my love for them both is immeasurable.

To all of you mentioned here above, and to many more who I could not fit within these pages, I owe a debt of gratitude that words cannot fully express. Your support, in its many forms, has been integral to my journey. Thank you for being part of my story.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In this chapter, I will provide an overview of the foundational topics that underpin this thesis. Section 1.1 delves into the cosmological background, offering an overview of the *Standard Model of Cosmology* and its general relativity foundations. It also elucidates several key concepts that are central to the thesis. Section 1.2 serves as an introduction to some of the primary techniques that cosmologists use to study the cosmos – galaxy clustering and weak gravitational lensing. This section outlines the formalism and elucidates the challenges associated with contemporary cosmological observations. As much of cosmological research relies on statistical techniques, Section 1.3 introduces a number of fundamental statistical concepts that will play a pivotal role throughout the entirety of the thesis. In recent years, Machine Learning has been emerging in cosmology as well. One of the key contributions of this thesis uses an unsupervised machine learning technique, so I will review some key concepts here as well. Section 1.4 provides an overview of the thesis structure, offering a roadmap for the subsequent chapters and their interconnections.

1.1 Standard Model of Cosmology

The Expanding Universe

In the early 20th century, astronomers found that the spectral lines in the light from distant galaxies were shifted towards the red end of the electromagnetic spectrum. This shift, known as *redshift*, was first observed by Vesto M. Slipher (1917), and plays a central role in this thesis. The redshift quantifies the relative change in a spectral line's wavelength compared to its original wavelength, denoted as λ_{emit}. It can be calculated using the formula:

$$z = \frac{\lambda_{obs}}{\lambda_{emit}} - 1. \tag{1.1}$$

While the intrinsic motion of a galaxy relative to the reference coordinate system can cause a redshift, commonly referred to as the Doppler redshift, Edwin Hubble compared the redshift of galaxies to their distance from Earth (Hubble, 1929), finding that galaxies farther away had larger

redshifts, which indicated that they were receding from us. This relationship is now known as Hubble's Law, and can be expressed as

$$v = H_0 d, \tag{1.2}$$

where v is the recessional velocity, d is the distance to the galaxy, and H_0 is the Hubble constant. The key insight here is that this observed redshift of galaxies is not due to any peculiar motion of the galaxies themselves, but rather represents a fundamental property of the universe itself. In fact, theoretical models by Alexander Friedman (1922) and Georges Lemaître (1927), based on Einstein's field equations in general relativity (Einstein, 1916), independently predicted an expanding universe. Friedmann showed that the universe could be described as expanding or contracting, and the scale factor $a(t)$ was introduced to quantify how the size of the universe changes with time. It is related to the cosmological redshift, z, as

$$z = \frac{a(t_0)}{a(t)} - 1, \tag{1.3}$$

where its value at the present age of the universe is defined to be $a(t_0) = 1$. It is also useful to define the Hubble rate in terms of the scale factor

$$H(t) \equiv \frac{1}{a}\frac{\mathrm{d}a}{\mathrm{d}t}. \tag{1.4}$$

Those works, alongside with the works of Howard Robertson (1929) and Arthur Walker (1937) played a pivotal role in establishing the expanding universe. They also formed an early theoretical basis for the *Cosmological Principle*, by describing the metric of a universe that is both homogeneous and isotropic at large scales. We will look into the Friedmann-Lemaitre-Robertson-Walker (FLRW) metric in more detail in Section 1.1.1.

Big Bang and the Cosmic Microwave Background

Building upon the earlier work of Lemaître, George Gamow (1946) made significant contributions to the Big Bang theory, particularly with his work on Big Bang Nucleosynthesis (BBN). He proposed that during the early moments of the universe, conditions were suitable for the fusion of light elements. Gamow and his collaborators, Ralph Alpher and Robert Herman, predicted the existence of residual radiation from the Big Bang, a sort of afterglow from the hot, early universe (Alpher, Herman & Gamow 1948). In 1965, the discovery of the Cosmic Microwave Background (CMB) Radiation by Arno Penzias and Robert Wilson was a monumental moment in cosmology (Penzias & Wilson 1965). This discovery provided strong empirical evidence for the Big Bang theory, as it was the predicted remnant radiation from the universe's hot, dense, early state. Subsequent observations confirmed the predictions of Big Bang Nucleosynthesis regarding the abundances of light elements (like hydrogen, helium, and lithium) in the universe (Peebles 1966;Yang et al. 1984). The Hot Big Bang model, which describes the universe as evolving from an extremely hot and dense initial state, not only explains the expansion of the universe but also

provides a coherent framework for understanding the formation of basic elements and the cosmic background radiation.

Despite its successes, the Big Bang Theory faced challenges, such as explaining the uniformity of the CMB temperature, also known as the horizon problem (how regions of the universe far beyond causal reach could have the same temperature), and the flatness problem (why the density of the universe is so close to the critical value required for a flat universe, despite sensitive dependence on initial conditions). These issues were later addressed with the introduction of the inflationary paradigm in the 1980s (Guth 1981; Linde 1982; Albrecht & Steinhardt 1982). Inflation proposes an exponential expansion in the universe's first moments, stretching space-time and evening out temperature differences, and driving the universe's density towards the critical value for flatness, regardless of initial conditions.

Dark Matter and Dark Energy

The latter part of the 20th century in cosmology was marked by two monumental discoveries that profoundly changed our understanding of the universe: dark matter and dark energy. These components are not directly observable but are inferred from their gravitational effects and the expansion of the universe, respectively. The first galaxy surveys also started to provided evidence of inhomogeneity and anisotropy in the distribution of galaxies at small-scales (Geller & Huchra 1989; York et al. 2000; Colless et al. 2001).

Vera Rubin and Kent Ford's observations of the rotation curves of galaxies provided crucial evidence for dark matter (Rubin & Ford, 1970). Their observation that stars in galaxies rotate at such speeds that they could not be held together by the gravitational pull of the visible matter alone, suggested the presence of an unseen mass, or dark matter. Observations of gravitational lensing (Tyson et al., 1984), where the light from distant galaxies is bent by the gravitational field of a foreground object, further supported the existence of dark matter. In addition, more detailed studies of the CMB by Bennett et al. (1994) provided indirect evidence for dark matter through the analysis of the temperature fluctuations in the CMB. These fluctuations reflect density variations in the early universe, which align more closely with theories including dark matter.

In 1998, two independent research teams, the High-Z Supernova Search Team (Riess et al., 1998) and the Supernova Cosmology Project (Perlmutter et al., 1999), made an astounding discovery: the expansion rate of the universe is accelerating. They used Type Ia supernovae as standard candles to measure the expansion rate and found that distant supernovae were dimmer than expected, indicating that the universe was expanding more rapidly than in the past. This acceleration implied the existence of an unknown energy component, now known as dark energy, working against the pull of gravity and responsible for the accelerated expansion. Einstein's cosmological constant Λ, once introduced to allow a static universe and later abandoned (Einstein, 1916), gained new significance as a plausible candidate for dark energy.

ΛCDM and the Era of Precision Cosmology

The 21st century, often described as era of *precision cosmology*, is characterized by unprecedented advancements in observational techniques, data analysis, robust theoretical frameworks, and

international collaboration. The *Planck* satellite has provided incredibly detailed observations of the cosmic microwave background (CMB) and the early universe (Planck Collaboration et al., 2016, 2018). Large-scale survey projects, including the first ones like the Sloan Digital Sky Survey (SDSS) (York et al., 2000), Wilkinson Microwave Anisotropy Probe (WMAP) (Bennett et al., 2003), the current Stage III surveys like the Kilo-Degree Survey (KiDS) (de Jong et al., 2013), the Hyper Suprime-Cam survey (HSC) Aihara et al. 2018; Hamana et al. 2020, and the Dark Energy Survey (DES) DES Collaboration 2016a; Troxel et al. 2018, and the forthcoming Legacy Survey of Space and Time (LSST) by the Vera C. Rubin Observatory (Ivezić et al., 2019) Nancy Grace Roman Space Telescope (Spergel et al., 2015), and Euclid (Laureijs et al., 2011), are mapping millions of galaxies, providing a comprehensive view of the universe's large-scale structure.

The standard model of cosmology, often referred to as the Lambda Cold Dark Matter (ΛCDM) model, is the prevailing framework used to describe the properties and evolution of the universe. It employs Einstein's general relativity to characterize the gravitational phenomena of the universe and posits a large-scale uniformity and isotropy, also incorporating the observed expansion of the universe. In addition, it integrates observations from the several probes described above to elucidate the expansion history of the universe. In this model, dark energy, in the form of the cosmological constant, drives the accelerated expansion of the Universe, while cold dark matter plays a major role in explaining the formation and the clumping of structure.

The ΛCDM model continues to be tested against increasingly precise data. The nature of dark energy and dark matter remains a central focus. Precision cosmology aims to resolve current puzzles like the apparent Hubble tension: discrepancy between the current rate of expansion of the universe as predicted by CMB measurements, which are based on the early universe under the ΛCDM model, and the rate derived from direct measurements in the local universe (see Valentino et al. (2021) for a review, and also Riess et al. (2019); Verde et al. (2019)); and the possible σ_8 tension: the conflict between the amount of clustering of matter in the universe as predicted by ΛCDM, when calibrated with CMB observations, and observations in the late-time universe through galaxy surveys and other structure measurements (Dark Energy Survey Collaboration & Kilo-Degree Survey Collaboration, 2023; Miyatake et al., 2023; Sugiyama et al., 2023; DES Collaboration, 2018; Asgari et al., 2021). The resolution of these apparent tensions is crucial for enhancing our understanding of the fundamental constituents of the Universe and may point the way to a new paradigm in cosmology.

1.1.1 Framework and Parameterization

In this subsection I provide a summarized introduction to the theoretical framework and the components of the cosmological model we will use throughout this thesis. I refer to Dodelson & Schmidt (2020) for a detailed read on the topics presented here.

Cosmology uses Albert Einstein's theory of general relativity as the basis for describing the gravitational behavior of the universe. The Friedmann-Lemaître-Robertson-Walker (FLRW) metric, $g_{\mu\nu}$, for a Euclidean, or flat, universe is given by

$$\mathrm{d}s^2 = g_{\mu\nu}\mathrm{d}x^\mu \mathrm{d}x^\nu = -\mathrm{d}t^2 + a^2(t)\delta_{ij}\mathrm{d}x^i \mathrm{d}x^j \tag{1.5}$$

where I am using Einstein's summation notation, and the units where $\hbar = c = k_B = 1$. Applying the field equations of General Relativity to the FLRW metric, we can derive the Friedmann equations, which relates the evolution of the scale factor with the composition of the homogeneous constituents in the universe:

$$H^2(t) = \left(\frac{\dot{a}(t)}{a(t)}\right)^2 = \frac{8\pi G}{3}\rho(t), \tag{1.6}$$

and

$$\frac{\ddot{a}(t)}{\dot{a}(t)} = -\frac{4\pi G}{3}\left[\rho(t) + 3p(t)\right], \tag{1.7}$$

where G is Newton's constant, $\rho(t)$ is the energy density in the universe as a function of time, and $p(t)$ is the pressure.

Distances

The comoving coordinate system, is a spatial coordinate system that is stationary relative to the large-scale motion of the universe. This is a convenient system that allows us to think of the physical distance between us and an object in an expanding universe as the comoving distance scaled by the scale factor:

$$d(\chi, t) = a(t)\chi. \tag{1.8}$$

It can be written in terms of the scale factor $a(t)$ and the redshift z as:

$$\chi(t) = \int_t^{t_0} \frac{\mathrm{d}t'}{a(t')} = \int_{a(t)}^{1} \frac{\mathrm{d}a'}{a'^2 H(a')} = \int_0^z \frac{\mathrm{d}z'}{H(z')}. \tag{1.9}$$

Measurements of the cosmic expansion are generally done in terms of two types of cosmological distances, the angular diameter distance and the luminosity distance. The former relates the apparent angular size of an object in radians, θ, to its physical size, D:

$$d_A = \frac{D}{\theta}. \tag{1.10}$$

In ΛCDM, with Euclidean geometry, we have $\theta = (D/a)\chi(a)$, and therefore we can re-write

$$d_A^{\mathrm{Euc}} = a\chi = \frac{\chi}{1+z}. \tag{1.11}$$

The luminosity distance, relates the observed flux (integrated over all frequencies), F, of an object to its intrinsic luminosity, L, emitted in its rest frame:

$$d_L = \sqrt{\frac{L}{4\pi F}}; \qquad\qquad d_L^{\rm Euc} \equiv \frac{\chi}{a}. \tag{1.12}$$

The luminosity and angular diameter distances are related by

$$d_L = (1+z)^2 d_A. \tag{1.13}$$

The cosmological redshift, angular diameter distance and luminosity distance are the three basic distance metrics used in the field of observational cosmology.

Components

The ΛCDM universe contains three basic density components: radiation, matter, and dark energy. The radiation term includes all highly relativistic particles, like photons, which travel at or near the speed of light and are significant in the early universe. The matter term is divided into baryonic matter and dark matter. In a cosmological context, baryons refers collectively to all known forms of matter that emit or absorb photons, including protons, electrons and even neutrons that make up the visible components of galaxies. Dark matter is a phenomenological term to describe that mass component that must be present to explain gravitational effects observed but which cannot be accounted for by baryons. Dark energy accounts for the observed fact that the expansion of the universe is speeding up over time, contrary to the expectations from gravitational attraction alone.

On cosmological scales ($\gtrsim 10$ Mpc) each of these is well approximated as a continuous ideal fluid, which is governed by a linear equation of state relating the pressure p and energy density ρ, in the form

$$p_i = w_i \rho_i \tag{1.14}$$

where the subscript i denotes the component, and w_i is the dimensionless state parameter. Applying equation 1.6 and 1.7, the redshift evolution can be derived as,

$$\rho_i(z) = \rho_{i,0}(1+z)^{3(1+w_i)}, \tag{1.15}$$

where $\rho_{i,0} = \rho_i(z=0)$ is the present-day mean density of component i. Let us define the present-day density parameter of component i in units of the critical density

$$\Omega_i \equiv \frac{\rho_{i,0}}{\rho_{\rm cr}}, \tag{1.16}$$

where $\rho_{\rm cr} \equiv 3H_0^2/8\pi G$ is defined as the threshold density, below which a universe with no dark energy expands forever. Many observable probes are sensitive to physical rather than comoving densities, which mean it is often convenient to cast the density parameter in a form that depends on the expansion rate, $\Omega_i h^2$. The factor h represents the Hubble parameter in units of 100 km/s/Mpc,

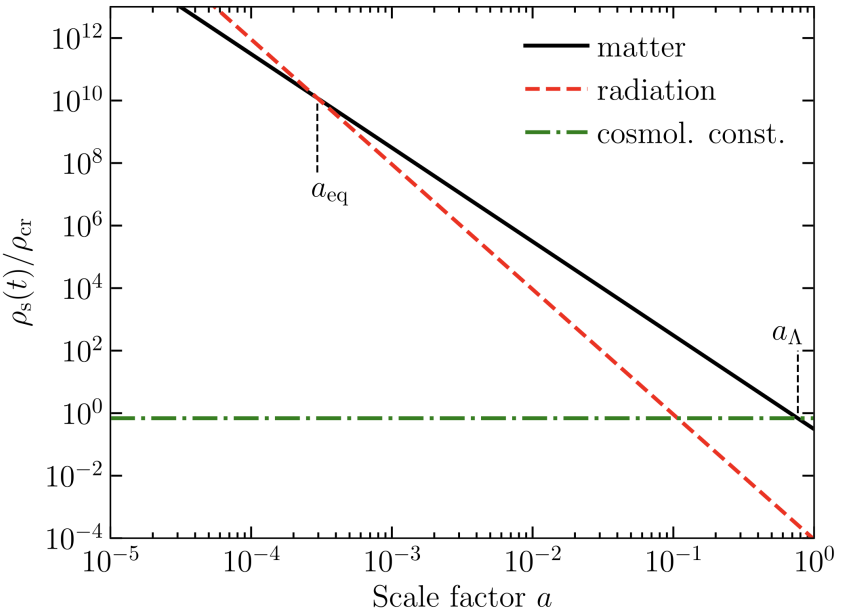

Figure 1.1: Taken from Dodelson & Schmidt (2020): Energy density as a function of scale factor for different constituents of the Euclidean fiducial cosmology. Matter ($\propto a^{-3}$), radiation ($\propto a^{-4}$), and a cosmological constant. All are in units of the critical density today. Even though matter and the cosmological constant appear to dominate today, at early times, the radiation density was largest. The epoch at which the energy densities of matter and radiation are equal is a_{eq}, while the epoch at which the densities of matter and cosmological constant match is a_Λ.

which accounts for the uncertainty in the actual value of the Hubble constant and allows for a way to express results that are somewhat independent of this uncertainty. We can then re-write the Friedman's equation in a more convenient form:

$$H(z) = H_0[\Omega_{\rm m}(1+z)^{-3} + \Omega_{\rm r}(1+z)^{-4} + \Omega_K(1+z)^{-2} + \Omega_{\rm de}(z)]^{1/2}. \tag{1.17}$$

The expansion is initially driven by radiation, with the universe gradually evolving through matter and finally dark energy dominated eras. Figure 1.1 shows the type of density energy that dominates in different epochs of the universe, as a function of the scale factor. Setting $t = t_0$, equation 1.17 yields $1 - \Omega_K = \Omega_{\rm m} + \Omega_{\rm r} + \Omega_{\rm de}$, where Ω_K is the curvature density. $\Omega_K > 0$ implies a spatially finite closed universe, which will recollapse; $\Omega_K < 0$ indicates an infinite open universe, which will expand forever; and $\Omega_K = 0$ implies an infinite, perpetually expanding universe with flat geometry. The latter is favored by all observational evidence, therefore a flat universe is what we have been assuming and will continue to assume in what follows.

1.1.2 Power Spectrum

Let us begin with a general treatment of the 2-point statistics of a generic density field $\rho(\vec{x})$. The density contrast of the field can be defined as

$$\delta(\vec{x}) \equiv \frac{\rho(\vec{x}) - \bar{\rho}}{\bar{\rho}}, \tag{1.18}$$

where $\bar{\rho}$ is the mean value of the density field. Then, by definition, the mean value of $\delta(\vec{x})$ is

$$\langle \delta(\vec{x}) \rangle = 0, \tag{1.19}$$

where the brackets stand for the average over the entire space. The next step in describing the statistics of the density fields is to consider the two-point correlation function,

$$\xi(\vec{x}, \vec{y}) \equiv \langle \delta(\vec{x}) \delta(\vec{y}) \rangle, \tag{1.20}$$

which encodes the information about statistical dependence of the density field between any two points.

We are going to assume the field to be homogeneously and isotropically distributed, which implies there is no preferred position or direction. As a consequence, the correlation function will depend only on the magnitude of the difference between $\vec{x}$ and $\vec{y}$,

$$\xi(\vec{x}, \vec{y}) \overset{\text{homogeneity}}{=} \xi(\vec{x} - \vec{y}) \overset{\text{isotropy}}{=} \xi(|\vec{x} - \vec{y}|). \tag{1.21}$$

While the statistical interpretation of the correlation function is quite intuitive, it is often useful to consider its Fourier transform, leading to the definition of the so-called power spectrum, $P(k)$,

$$\begin{aligned}
\langle \tilde{\delta}(\vec{k}) \tilde{\delta}(\vec{k}') \rangle &= \tilde{\xi}(\vec{k}, \vec{k}') = \int \mathrm{d}^3 x e^{-i\vec{k}\cdot\vec{x}} \int \mathrm{d}^3 y e^{-i\vec{k}'\cdot\vec{y}} \xi(\vec{x}, \vec{y}) \\
&= \int \mathrm{d}^3 x e^{-i\vec{k}\cdot\vec{x}} \int \mathrm{d}^3 y e^{-i\vec{k}'\cdot\vec{y}} \xi(\vec{x} - \vec{y}) \\
&= \int \mathrm{d}^3 x e^{-i(\vec{k}+\vec{k}')\cdot\vec{x}} \int \mathrm{d}^3 z e^{-i\vec{k}'\cdot\vec{z}} \xi(\vec{z}), \qquad (\ \vec{z} \equiv \vec{x} - \vec{y}\ ) \\
&= (2\pi)^3 \delta_{\mathrm{D}}^3(\vec{k} + \vec{k}') P(k),
\end{aligned} \tag{1.22}$$

where we have used homogeneity in the second line and

$$P(\vec{k}') = \int \mathrm{d}^3 z e^{-i\vec{k}'\cdot\vec{z}} \xi(\vec{z}) = P(|\vec{k}'|) \equiv P(k'). \tag{1.23}$$

Note that the dependence only in the magnitude of $\vec{k}'$ is a manifestation of isotropy in ξ and that we are allowed to exchange $P(k')$ by $P(k)$ in (1.22) due to the Dirac delta function, δ_{D}. Specifically, since the correlation function for the density field depends only on the distance between two points, the fluctuations in the Fourier modes are uncorrelated unless the two modes have equal

and opposite wave vectors, which is the information encoded in the delta function. Alternatively, one can also integrate (1.22) to write the power spectrum as

$$P(k) = \int \frac{\mathrm{d}^3 k'}{(2\pi)^3} \langle \tilde{\delta}(\vec{k}) \tilde{\delta}(\vec{k}') \rangle. \tag{1.24}$$

The power spectrum is a crucial statistical tool in cosmology, with a simple physical interpretation. We can think of the random field of mass density fluctuations in the universe at a given time as a collection of Fourier modes, or a superposition of standing waves with varying wavelengths. In this context, each k mode within the matter power spectrum corresponds to a sine wave with a specific angular frequency, and the power $P(k)$ represents the amplitude squared of that wave.

Post-inflation, it is widely held that the power spectrum of density perturbations adheres to a power-law distribution with an index n_s approximately equal to 1. This linear shape is called the Harrison-Zel'dovich-Peebles spectrum, and can be written as

$$P_{pr}(k, a) = A_s T(k, a) k^{n_s}, \tag{1.25}$$

where A_s is the amplitude of P_{pr}, and $T(k, a)$ is the transfer function, describing later linear modifications to the power spectrum. Before matter-radiation equality, Fourier modes with $\lambda < d_H$ are suppressed due to the dominance of radiation pressure, where d_H is the horizon distance. This defines a threshold k_t above which growth is suppressed. As d_H expands, k_t shifts downwards, and larger overdensities are gradually allowed to begin growing. The position of the peak is frozen at the time of equality, $k_t(a_{eq}) = k_0$, and subsequent growth, governed by the growth function $D_+(a)$, distributes power evenly across all k modes.

The growth function describes how these initial perturbations generated in the radiation-dominated epoch are amplified under the influence of gravity, when the universe transitions to a matter-dominated phase

$$P(k, a) = P_{pr}(k, a) D_+^2(a), \tag{1.26}$$

where $P(k, a)$ is the linear power spectrum of matter at late times.

With the onset of dark energy dominance (Λ in Figure 1.1), an additional complexity is introduced to the evolution of cosmic structure. Dark energy begins to oppose the gravitational attraction that drives the growth of structures. This interplay is captured by the growth function, which shows a modified growth rate in the presence of dark energy. Therefore, the growth function is sensitive to the overall matter density of the universe and the properties of dark energy, especially its equation of state w.

The matter power spectrum today has the approximate form

$$P(k) \propto \begin{cases} k & k < k_0 \\ k^{-3} & k > k_0 \, . \end{cases} \tag{1.27}$$

Figure 1.2 shows the matter power spectrum at redshift $z = 0$ in the fiducial ΛCDM cosmology for different values of Ω_m. The shape of $P(k)$ depends on the primordial spectrum and the peak position indicates d_H at equality, making it sensitive to $\Omega_\mathrm{m} h$. The amplitude of the linear spectrum

at a redshift of zero is set by a normalisation parameter σ_8, which by convention is defined as the root mean square (rms) of overdensity fluctuations, averaged in spheres of comoving radius $8h^{-1}$ Mpc. This is a key physical parameter describing the large scale clustering of the late-time universe, and can be measured directly by any data set which is sensitive to $P(k)$.

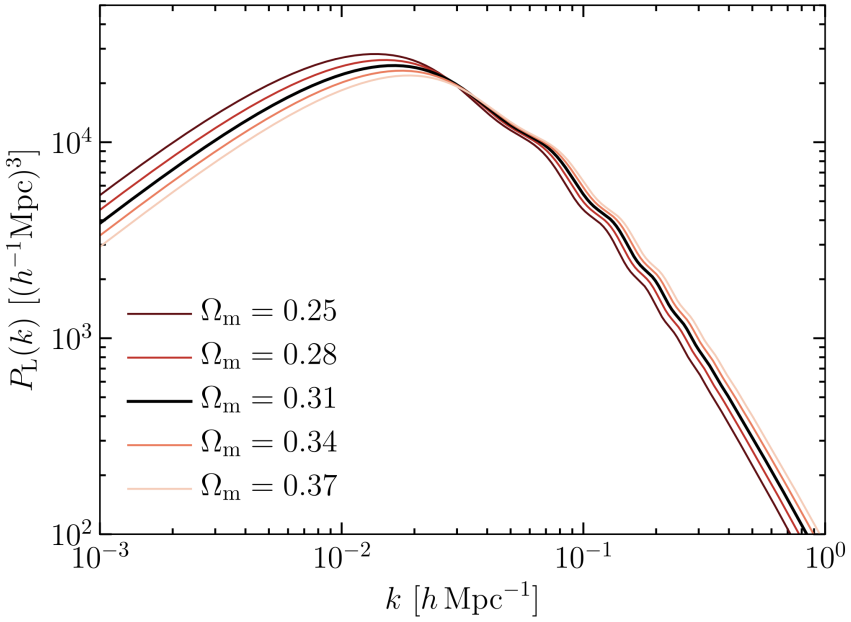

Figure 1.2: Taken from Dodelson & Schmidt (2020): The matter power spectrum at redshift $z = 0$ in the fiducial ΛCDM cosmology (thick black line). The other lines show the result when varying $\Omega_{\rm m}$ around the fiducial value, keeping h fixed and $\Omega_{\rm m} + \Omega_\Lambda = 1$. Changing $\Omega_{\rm m}$ changes the epoch of equality, k_0, and hence the shape of the matter power spectrum.

Theory makes no a priori predictions for the exact values of A_s, n_s, σ_8 beyond ruling out unphysical regions of parameter space. Thus, they must be treated as free parameters in the cosmological model, to be constrained by observation. The way in which changes in the linear fluctuation amplitude σ_8 affect $P(k)$ is strongly degenerate with variations in the background average mass density $\Omega_{\rm m}$, and it is thus common to define best-constrained combinations for particular observational probes (e.g. $S_8 = \sigma_8\sqrt{\Omega_{\rm m}/0.3}$ in cosmic shear).

1.2 Observational Cosmology

Observations of the large-scale distribution of galaxies allow us to study the large-scale structure of the universe. Wide-field surveys, like the the ongoing Kilo-Degree Survey(KiDS) (de Jong et al., 2013), the Dark Energy Survey (DES) (DES Collaboration, 2016a), and the Hyper Suprime-Cam survey (HSC) (Aihara et al., 2018), have mapped millions of galaxies, providing important insights into the properties of dark energy and the growth of structure.

In this section, I will review the basic theory of galaxy clustering and weak lensing – two pivotal observables in the analysis of large-scale structure. Additionally, the role and impact of photometric redshifts on these observables will be explored, alongside with the statistical treatment of their two-point correlations: galaxy-galaxy, shear-shear, and the cross-correlation of galaxy-shear. This section is mainly inspired and based on Dodelson & Schmidt (2020), Mandelbaum (2018), Dodelson (2017), Liddle (2015), and I refer to these works for a more complete read on the topics presented here.

1.2.1 Photometric Redshifts

Photometric redshift estimation is a critical piece in wide-field surveys analysis. These surveys have the advantage of a great sky area coverage in a relatively short time, which compensates for the lack of radial precision. The radial distances of galaxies can be estimated using *photometric* redshifts, by making observations in multi-band color filters. The redshifts estimated using filters are much less accurate than spectroscopic estimates, which measure the full spectra of light emitted for each object separately. Spectroscopic surveys provide high-precision radial distances; however, even though there are large spectroscopic surveys such as BOSS (Dawson et al., 2013) and DESI (Levi et al., 2013; DESI Collaboration, 2023), they are very resource-intensive and therefore observe far fewer galaxies.

The statistical power inherent in the large number of galaxies observed in photometric surveys helps to mitigate the lower precision in the redshift estimation. In these surveys, galaxies are sorted into redshift bins, a process analogous to tomographic imaging due to the stratified analysis of data layers. These bins reflect the mean redshift estimate and its uncertainty. This limitation has to be taken into account during the theoretical modeling such that the radial integrations of the correlation functions are done in the redshift bins, and the redshift uncertainty is accounted as a systematic error.

When modeling, we introduce a free parameter Δz (along with a prior) that accounts for the redshift uncertainty for bin i:

$$n^i(z) = n^i_{\rm PZ}(z - \Delta z^i), \tag{1.28}$$

where $n^i_{\rm PZ}$ is the estimated redshift distribution for the galaxy sample under consideration. This adjustment is crucial to accommodate the inherent uncertainties in the photometric redshift estimates. The redshift distribution of galaxies within a specific bin, i, is expressed as $n^i_{\rm g}(z)$. The mean angular number density for bin i is then given by the integral over the redshift distribution:

$$\bar{n}^i_{\rm g} = \int {\rm d}z\, n^i_{\rm g}(z). \tag{1.29}$$

Chapter 2 will delve into the details of photometric redshift characterization for weak lensing source galaxies, in the context of the method used in the Dark Energy Survey.

1.2.2 Galaxy clustering

Galaxy clustering has long served as a pivotal cosmological probe, with its foundations tracing back to the initial observations of complex patterns in the spatial distribution of galaxies across the sky. The study of these patterns has advanced significantly over the years, informed by sophisticated statistical analyses and the advent of large-scale surveys that provide comprehensive samples of the universe.

Angular correlation function and angular power spectrum

In the absence of full redshift information, the use of the three-dimensional power spectrum, detailed in Section 1.1.2, is not feasible. Photometric surveys provide two-dimensional positions for numerous galaxies but lack distance measurements. They do, however, collect multi-band photometric information that facilitates the estimation of photometric redshifts. This enables the use of the projected two-dimensional angular power spectrum to study of the cosmic structure.

Consider the projection of the galaxy density fluctuation $\delta_g(\vec{n}, z)$, where z is the redshift in a given direction $\vec{n}$ in the sky plane. Geometrically, we are projecting $\vec{x} = (x_1, x_2, x_3)$ in the (x_1, x_2)-plane and locating a point on the plane by a two-dimensional vector $\vec{n} = (\theta_1, \theta_2)$. This is called the flat-sky approximation. The projection of $\delta_g(\vec{n}, z)$ along some direction $\vec{n}$ in the sky plane is accomplished as

$$\delta(\vec{n}) = \int_0^\infty \mathrm{d}z\, W(z)\delta_g(\vec{n}, z), \tag{1.30}$$

where W is the radial selection function, which is the probability of observing a galaxy at redshift z, such that

$$\int_0^\infty \mathrm{d}z\, W(z) = 1. \tag{1.31}$$

One can also write the selection function in terms of the redshift distribution of galaxies, n_g, as follows,

$$W(z) = \frac{n_g(z)}{\bar{n}_g}. \tag{1.32}$$

The angular number density of galaxies in the redshift bin i is independent of whether the bin is parameterized by z of χ, meaning that

$$\mathrm{d}\bar{n}^i_{\mathrm{g}} = n^i_{\mathrm{g}}(z)\mathrm{d}z = n^i_g(\chi)\mathrm{d}\chi, \tag{1.33}$$

where $n^i_{\mathrm{g}}(z)$ is redshift distribution of the galaxy sample in this redshift bin. Therefore we can also express W in terms of the coming distance as:

$$W(\chi) = \int \mathrm{d}\chi' n(\chi'). \tag{1.34}$$

As we have seen in Section 1.1.2, the power spectrum is the Fourier transform of the correlation function. In this section, we denote the $2D$ power spectrum by $C^i_{\delta_g\delta_g}(\ell)$, such that

$$\begin{aligned} w^i(\theta) &= \int \frac{\mathrm{d}^2\ell}{(2\pi)^2} e^{i\vec{\ell}\cdot\vec{\theta}} C^i_{\delta_g\delta_g}(\ell) = \int_0^\infty \frac{\mathrm{d}\ell\,\ell}{2\pi} C^i_{\delta_g\delta_g}(\ell) \left(\frac{1}{2\pi}\int_0^{2\pi} \mathrm{d}\phi e^{i\ell\theta\cos\phi}\right) \\ &= \int_0^\infty \frac{\mathrm{d}\ell\,\ell}{2\pi} J_0(\ell\theta) C^i_{\delta_g\delta_g}(\ell). \end{aligned} \tag{1.35}$$

The $2D$ power spectrum can be expressed in terms of the $3D$ one as

$$C^i_{\delta_g\delta_g}(\ell) = \int_0^{\chi_\infty} d\chi \frac{W^i_{\delta_g}(\chi)^2}{\chi^2} P_\delta\left(\frac{\ell+1/2}{\chi}, z(\chi)\right). \tag{1.36}$$

Equations 1.35 and 1.36 allow us to utilize the two-dimensional statistics on galaxy clustering from photometric surveys to infer the three-dimensional power spectrum and hence get insight in the underlying mass distribution in the universe.

Galaxy bias

Galaxies are biased tracers of the underlying total matter field. This concept was introduced by Kaiser (1984) through the idea that galaxies are quite rare objects, standing on the peaks in the matter distribution. The effect of this hypothesis in the correlation function of galaxies is an increment relative to the mass by the so-called bias parameter, b:

$$\delta_g(x) = b\delta_m(x), \tag{1.37}$$

linking relative fluctuations in the galaxy field δ_g to those in the matter field δ_m. The bias factor is not directly measurable but must be inferred from observations, such as galaxy clustering statistics, and can vary with galaxy properties such as luminosity, color, and morphology. At large scales, we can safely assume a linear bias model, with a single parameter for each redshift bin i:

$$b^i(z) = b^i_1, \tag{1.38}$$

where the subscript "1" represents the linear term in the Taylor expansion of the galaxy density field with respect to the matter density field. Then, the radial weight function for angular galaxy clustering can be expressed as:

$$W^i_{\delta_g}(\chi) = b^i_1 \frac{n^i_g(z(\chi))}{\bar{n}^i_g} \frac{\mathrm{d}z}{\mathrm{d}\chi}. \tag{1.39}$$

On smaller scales, particularly those comparable to the sizes of galaxies or galaxy groups, non-linear, stochastic, and scale-dependent aspects of biasing come into play. Galaxy bias is determined by the physics of galaxy formation and evolution, which are influenced by the local environment. Therefore, it is a function of both the scale and the redshift, with galaxies at higher

redshifts exhibiting a stronger bias at small scales due to the nature of structure formation in the early universe.

In cosmological analyses, galaxy bias is a nuisance parameter that must be marginalized over or otherwise accounted for to extract precise measurements of cosmological parameters. If not accounted for, it can mimic or mask the signatures of other effects, such as those from dark energy.

1.2.3 Weak lensing

Gravitational lensing, the deflection of light by matter as it traverses the cosmos, is a pivotal phenomenon in observational cosmology. In its strong form, it can produce dramatic effects such as the formation of multiple images of a single astronomical source. However, it is the subtle regime of this process, known as *weak lensing*, that offers invaluable insights for cosmology. Crucially, weak lensing serves as an unbiased probe of the total matter distribution, including dark matter, independent of the complex galaxy bias that affects other observational methods, such as galaxy clustering as we discussed in the previous section. By directly mapping the mass distribution in the universe irrespective of the luminous matter, weak lensing allows us study the cosmic structure and the dark sector, providing essential insights into the fundamental constituents and evolution of the universe.

Lens equation

Consider a light ray traveling from a faraway source to an observer, separated by a distance $D_{\rm S}$, that passes near to a gravitational potential – represented by the shaded region in Figure 1.3 – which acts as a lens, at a distance $D_{\rm LS}$ from the source. We also define the line connecting the observer to the lens as the z-axis and $D_{\rm L} = D_{\rm S} - D_{\rm SL}$ to be the distance between the lens and the observer. There are three relevant angles that can be thought as vectors in the xy-plane containing the sheared image: the true angular position of the source, $\vec{\beta}$, the position where the observer sees the light coming from due the lensing, $\vec{\theta}$, and the deflection angle, $\vec{\alpha}$.

From the relation between those three angles we can write the *lens equation*,

$$\vec{\beta} = \vec{\theta} - \vec{\alpha}(\vec{\theta}), \tag{1.40}$$

which reflects the fact that on its journey from the source up to the observer, light encounters several lensings, undergoing multiple deflections due to the variation of the gravitational potential, ϕ, along the line of sight. Let us define the projected gravitational potential

$$\Phi(\vec{\theta}) \equiv \frac{2}{D_{\rm S}} \int_0^{D_{\rm S}} {\rm d}D_{\rm L} \frac{D_{\rm SL}}{D_{\rm L}} \phi(D_{\rm L}\theta^i, D_{\rm L}; t_0 - D_{\rm L}/c), \tag{1.41}$$

using the geodesic equation for the emitted light, one can show that

$$\alpha^i(\vec{\theta}) = \frac{1}{c^2} \frac{\partial}{\partial \theta^i} \Phi(\vec{\theta}), \tag{1.42}$$

where α^i are the components of the $2D$ deflection angle.

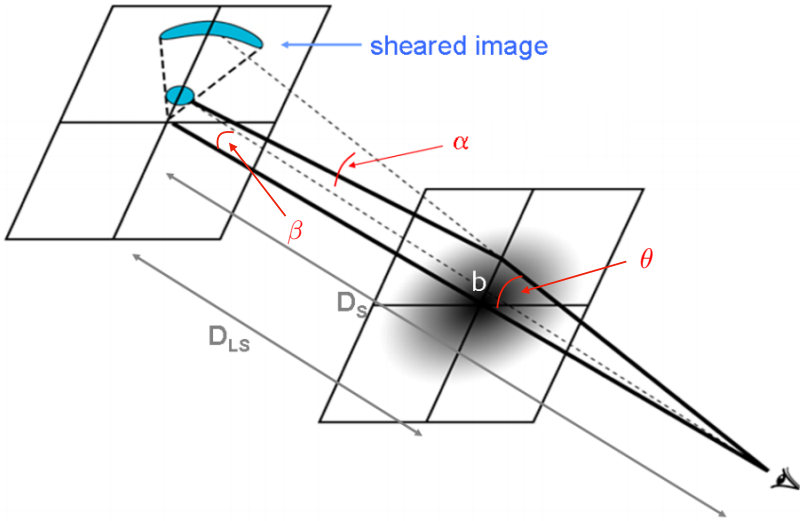

Figure 1.3: Figure adapted from Mandelbaum (2018).

Distortion tensor

The surface brightness, $\mathcal{S}$, of an object or region in the sky is defined as the flux (energy per unity of time per unity of area) per solid angle. By definition, $\mathcal{S}$ remains constant no matter how far the object is from the observer: the drop in the flux as the object moves away is compensated by the increase in the physical size subtended by a square arcsecond. This conservation of surface brightness is expressed as $\mathcal{S}^{\text{un-lensed}}(\vec{\beta}) = \mathcal{S}^{\text{lensed}}(\vec{\theta})$.

The magnification μ is the ratio of the lensed to the unlensed flux,

$$\mu \equiv \frac{\mathcal{S}^{\text{lensed}}(\vec{\theta})\mathrm{d}^2\theta}{\mathcal{S}^{\text{un-lensed}}(\vec{\beta})\mathrm{d}^2\beta} = \left|\frac{\mathrm{d}^2\theta}{\mathrm{d}\beta^2}\right|, \tag{1.43}$$

which is the Jacobian of the transformation from the source area $\mathrm{d}^2\beta$ to the image area $\mathrm{d}^2\theta$. Using the lens equation (1.40), the Jacobian matrix acquires the form

$$\frac{\partial\beta_i}{\partial\theta_j} = \begin{pmatrix} 1 - \frac{\partial\alpha_x}{\partial\theta_x} & -\frac{\partial\alpha_x}{\partial\theta_y} \\ -\frac{\partial\alpha_y}{\partial\theta_x} & 1 - \frac{\partial\alpha_y}{\partial\theta_y} \end{pmatrix} \equiv \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \Psi_{ij}. \tag{1.44}$$

Note that the off-diagonal terms are equal due to (1.42). Therefore, we are left with three independent terms describing the effects of lensing: the convergence, κ, and the two components of shear, γ_1 and γ_2, composing the so-called distortion tensor, Ψ_{ij},

$$\Psi_{ij} \equiv \begin{pmatrix} \kappa + \gamma_1 & \gamma_2 \\ \gamma_2 & \kappa - \gamma_1 \end{pmatrix}, \tag{1.45}$$

quantifying the deviation of the Jacobian matrix from unity. Figure 1.4 provides an illustration of the effect of each of its components. Alternatively, its definition in terms of the projected potential corresponds to

$$\kappa \equiv \frac{1}{2c^2}\left(\frac{\partial^2}{\partial\theta_x^2}+\frac{\partial^2}{\partial\theta_y^2}\right)\Phi, \qquad \gamma_1 \equiv \frac{1}{2c^2}\left(\frac{\partial^2}{\partial\theta_x^2}-\frac{\partial^2}{\partial\theta_y^2}\right)\Phi, \qquad \gamma_2 \equiv \frac{1}{c^2}\frac{\partial^2\Phi}{\partial\theta_x\partial\theta_y}. \tag{1.46}$$

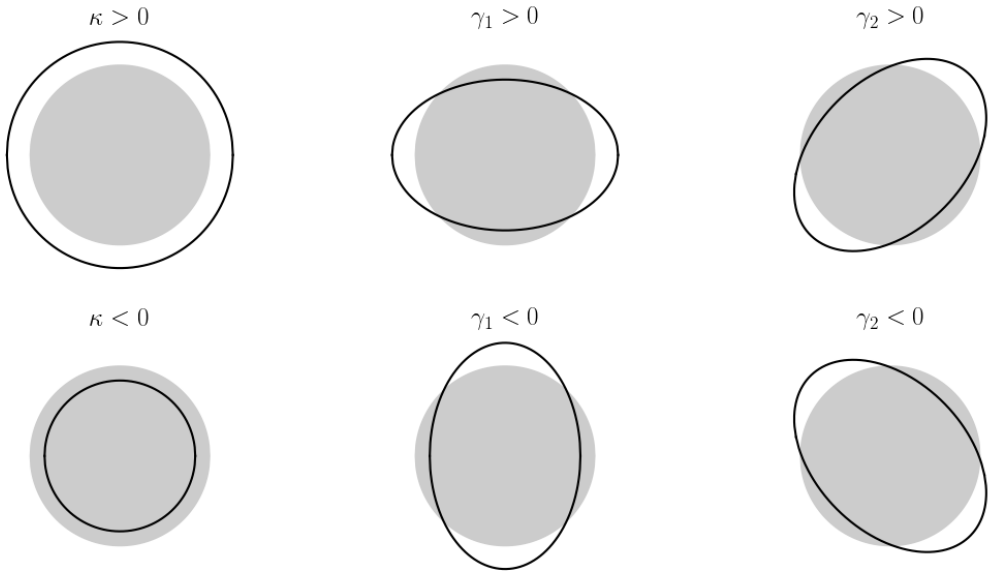

Figure 1.4: Graphical illustration of convergence and shear. The shaded circles represent an un-lensed source, while the black ellipses are the same source under various types of distortion.

Convergence Power Spectrum

Convergence, κ, describes the magnification or demagnification of images due to lensing. It is directly related to the scalar potential of the mass distribution along the line of sight. The power spectrum of this convergence quantifies the variance of these convergence fields as a function of scale and hence is a direct probe of the matter power spectrum at different cosmic epochs. Let us substitute the projected gravitational potential, defined via (1.42), in (1.46),

$$\begin{aligned}\kappa(\vec{\theta}) &= \frac{1}{c^2}\int_0^{D_\mathrm{S}} \mathrm{d}D_\mathrm{L}\frac{D_\mathrm{SL}D_\mathrm{L}}{D_\mathrm{S}}\nabla^2\phi(x^i = D_\mathrm{L}\theta^i, D_\mathrm{L}) \\ &= \int_0^{D_\mathrm{S}} \mathrm{d}D_\mathrm{L} W(D_\mathrm{L}, D_\mathrm{S})\delta(D_\mathrm{L}\vec{\theta}, D_\mathrm{L}).\end{aligned} \tag{1.47}$$

The function W is usually referred to as the weighting function, since it weights the density along the line of sight. Note that it resembles the radial selection function (1.30) we encountered in

Section 1.2.2, justifying the same notation (though with a slightly different interpretation). In addition, the scale factor a is evaluated at a redshift corresponding to the distance $D_{\rm L}$. We can re-write Equation 1.47 in terms of the comoving distance and redshift as

$$\kappa(\vec{\theta}) = \int_0^{\chi_{\rm s}} {\rm d}\chi \, W_\kappa(\chi)\delta(\chi\vec{\theta}, \chi). \tag{1.48}$$

Now that we have the expression for the effect of lensing on a particular object, we can go ahead and compute the two point correlation function of κ between many lensed objects. In other words, we can write the $2D$ power spectrum of κ in the same way we obtained the one for δ_g

$$C^{ij}_{\kappa\kappa}(\ell) = \int_0^{\chi(z_{\rm max})} {\rm d}\chi \frac{W^i_\kappa(\chi) W^j_\kappa(\chi)}{\chi^2} P_\delta\left(\frac{\ell + 1/2}{\chi}, z(\chi)\right), \tag{1.49}$$

where, for each redshift bin i, we have

$$W^i_\kappa(\chi) = \frac{3H_0^2\Omega_{\rm m}}{2c^2}\frac{\chi}{a(\chi)}\int_\chi^{\chi_{\rm H}} {\rm d}\chi' \, n^i\left(z(\chi')\right)\frac{{\rm d}z}{{\rm d}\chi'}\frac{\chi' - \chi}{\chi'}. \tag{1.50}$$

The convergence power spectrum is related to the matter power spectrum via the redshift distribution of the ensemble of galaxies at redshift bin i and bin j, explicitly showing that uncertainty in the redshift distribution impacts the weak lensing analysis as well. Equipped with this power spectrum, we are now able to write the shear–shear and galaxy–shear correlation functions.

Cosmic Shear

Cosmic shear is a phenomenon that involves the slight distortions of the images of distant galaxies due to the bending of light by the gravitational field of matter. Shear-shear correlations are direct correlations of the shapes of galaxies. If the distortion of galaxy shapes is random, the correlation function should be zero at all separations. However, the presence of a cosmic shear signal would result in a non-zero correlation function, indicating that the shapes of galaxies are coherently distorted by intervening mass structures.

The two-point statistics that quantify correlations between the shapes of galaxies is captured by a pair of two-point functions:

$$\xi^{ij}_\pm = \langle\gamma^i_t\gamma^j_t\rangle_\theta \pm \langle\gamma^i_\times\gamma^j_\times\rangle_\theta, \tag{1.51}$$

defined for two populations of galaxies at redshift bins i and j, and with the angular brackets denoting averaging over galaxy pairs separated by a particular angular distance θ. The subscripts $(t, \times)$ indicate the orthogonal components of shear rotated into coordinate axes defined by the separation vector between the galaxies:

$$(\gamma_t, \gamma_\times) = (-(\gamma_1\cos(2\phi) + \gamma_2\sin(2\phi)), -(\gamma_2\cos(2\phi) - \gamma_1\sin(2\phi))), \tag{1.52}$$

where ϕ is the polar angle of the vector connecting the lens and the source galaxy position. By noting the equivalence of $\gamma = \gamma_t + i\gamma_\times = \gamma_1 + i\gamma_2$, and writing the shears in terms of their inverse Fourier transforms it can be shown that

$$\xi_+^{ij}(\theta) = \frac{1}{2\pi}\int \mathrm{d}\ell\, \ell J_0(\ell\theta) C_{\kappa\kappa}^{ij}(\ell) \tag{1.53}$$

$$\xi_-^{ij}(\theta) = \frac{1}{2\pi}\int \mathrm{d}\ell\, \ell J_4(\ell\theta) C_{\kappa\kappa}^{ij}(\ell) \tag{1.54}$$

where the Bessel function of the first kind,

$$J_n(x) = \frac{1}{i^n\pi}\int_0^\pi e^{ix\cos(\alpha)}\cos(n\alpha)\mathrm{d}\alpha, \tag{1.55}$$

serves as an effective window function onto the angular shear spectrum. The distortions due to cosmic shear are typically very small and not noticeable for individual galaxies without a comparison to the undistorted shape which is unknown. However, by statistically analyzing the shapes of many galaxies over a portion of the sky, we can detect a coherent pattern of alignment and distortion. Cosmic shear is particularly powerful because it is sensitive to all matter, not just the luminous matter we can see with telescopes. Therefore, it is a direct probe of the mass distribution, and is less susceptible to some of the biases, such as galaxy bias, that can affect other observables.

Galaxy-Galaxy Lensing

Position-shear correlations are correlations between the positions of galaxies and the shear of background galaxies (galaxy-galaxy lensing). The tangential shear is related to the projected mass distribution around the lens galaxy and is used to probe the dark matter halo properties of the lens.

The tangential shear is related to the excess surface density $\Delta\Sigma$ around the lens galaxy by:

$$\gamma_t(R)\Sigma_{\rm crit} = \Delta\Sigma(R) \equiv \bar{\Sigma}(<R) - \Sigma(R), \tag{1.56}$$

where $\bar{\Sigma}(<R)$ is the mean surface density inside the radius R, $\Sigma(R)$ is the surface density at radius R, and $\Sigma_{\rm crit}$ is the critical surface density defined as:

$$\Sigma_{\rm crit} = \frac{c^2}{4\pi G}\frac{D_S}{D_L D_{LS}}. \tag{1.57}$$

The galaxy–shear correlation function captures the characteristic distortion of the shapes of background source galaxies due the mass associated with foreground lenses. This is due to the tangential shear, which gives rise to a correlation between the density field of the lens with the distortion on the source via

$$\gamma_{\rm t}^{ij}(\theta) = \int \frac{\mathrm{d}\ell\,\ell}{2\pi} J_2(\ell\theta) C_{\delta_{\rm g}\kappa}^{ij}(\ell) = \int \frac{\mathrm{d}\ell\,\ell}{2\pi} J_2(\ell\theta)\int \mathrm{d}\chi \frac{W_{\delta_{\rm g}}^i(\chi) W_\kappa^j(\chi)}{\chi^2} P_\delta\left(\frac{\ell+1/2}{\chi}, z(\chi)\right).$$

By measuring the tangential shear as a function of radius from the lens galaxy, we can infer the mass profile of the lens galaxy's dark matter halo. This is particularly useful for understanding the properties of dark matter, such as its distribution and its interaction with baryonic matter.

Systematics

In cosmological analyses, a significant amount of effort is dedicated to understanding and mitigating systematics to ensure that the constraints on cosmological parameters are robust. Systematic errors can mimic real signals and lead to incorrect interpretations of the data. Unlike statistical errors, which are random and can be reduced by increasing the sample size, systematic errors can persist or even scale with the amount of data and need to be identified and corrected for to avoid misleading conclusions about our cosmological analysis. There are several source of systematics in large-scale structure analysis, both for galaxy clustering and weak lensing. I already touched upon two sources of systematics in the previous subsections, redshift uncertainties and galaxy bias, and in this subsection I will be introduction other two sources of systematics that enter in our analysis, shear calibration and intrinsic alignments.

Shear calibration: While measuring the shear effect, a multiplicative error can be generated by a variety of sources. For example, in ground-based telescopes the images of galaxies can be blurred due to atmospheric effects. To produce the observed image, a point-spread function (PSF) is convolved with the true image of the galaxy, and this process introduces a multiplicative error.

The multiplicative shear calibration is currently modeled using one parameter m^i per redshift bin, which affects cosmic shear and galaxy–galaxy lensing correlation functions, (1.54) and (1.58), respectively, via

$$\begin{aligned} \xi^{ij}_{+/-}(\theta) &\rightarrow (1+m^i)(1+m^j)\xi^{ij}_{+/-}(\theta) \\ \gamma_{\rm t}^{ij}(\theta) &\rightarrow (1+m^j)\gamma_{\rm t}^{ij}(\theta). \end{aligned} \tag{1.58}$$

We then marginalize over all four m^i independently, usually assuming Gaussian priors with mean 0 and width of about 0.01, which corresponds to a 1% uncertainty in the shear calibration in the most recent analysis.

Intrinsic alignment: Besides lensing, there are several other physical effects that could cause the alignment of galaxies shapes. Since we do not have access to the unlensed field, it is a quite complex problem to determine whether these alignments are due to shear effects or are intrinsic to the galaxy field. These *intrinsic alignments* (IA) are the major theoretical uncertainty for weak lensing (Mandelbaum, 2018). I will discuss IAs in detail in Chapter 3.

1.2.4 3×2-point correlations

Collecting the results from previous sections, we summarize the two-point correlation functions for galaxy clustering, cosmic shear and galaxy–galaxy lensing:

$$
\begin{aligned}
w^i(\theta) &= \int \frac{\mathrm{d}\ell\,\ell}{2\pi} J_0(\ell\theta) \int \mathrm{d}\chi \frac{W^i_{\delta_\mathrm{g}}\left(\frac{\ell+1/2}{\chi}, \chi\right) W^j_{\delta_\mathrm{g}}\left(\frac{\ell+1/2}{\chi}, \chi\right)}{\chi^2} P_\delta\left(\frac{\ell+1/2}{\chi}, z(\chi)\right) \\
\xi^{ij}_{+/-}(\theta) &= (1+m^i)(1+m^j) \int \frac{\mathrm{d}\ell\,\ell}{2\pi} J_{0/4}(\ell\theta) \int \mathrm{d}\chi \frac{W^i_\kappa(\chi) W^j_\kappa(\chi)}{\chi^2} P_\delta\left(\frac{\ell+1/2}{\chi}, z(\chi)\right) \\
\gamma^{ij}_\mathrm{t}(\theta) &= (1+m^i) \int \frac{\mathrm{d}\ell\,\ell}{2\pi} J_2(\ell\theta) \int \mathrm{d}\chi \frac{W^i_{\delta_\mathrm{g}}\left(\frac{\ell+1/2}{\chi}, \chi\right) W^j_\kappa(\chi)}{\chi^2} P_\delta\left(\frac{\ell+1/2}{\chi}, z(\chi)\right) \qquad (1.59)
\end{aligned}
$$

Equation 1.59 is the theory that depends on Ω_m and σ_8 through the power spectrum. Theory is compared to the data, which will be described in latter chapters. The comparison is done with Bayesian statistics, which I describe in the next section.

Figure 1.5 shows the cosmological parameter contours obtained with this observables, as well as with their combination, the 3×2-point correlation function for the Dark Energy Survey Year 3. We can see that the combined evaluation of these 3 two-point correlation function provides strong constraints in the Ω_m and σ_8 parameters. Since these two parameters are correlated, it is useful to introduce the $S_8 = \sigma_8\sqrt{\Omega_\mathrm{m}/0.3}$ parameter.

1.3 Statistical Methods and Machine Learning

Statistical analysis is an integral part of cosmology, and we rely on it to analyse and interpret vasts amounts of observational data. Statistical measures, such as the power spectrum and the two-point correlation function – referenced in the previous sections – are vital for quantifying the underlying properties of the cosmic structure and testing the predictions of different cosmological models. This section aims to review additional statistical tools that are integral to the analyses presented later in this thesis. In addition, this section will introduce some foundational concepts of Machine Learning and preview a specific technique that will be instrumental in Chapter 2.

1.3.1 Bayesian Statistics

Bayesian statistics is grounded in a distinctive interpretation of probability as a measure of certainty. Traditionally, probability is viewed as the long-run frequency of an event occurring after numerous trials. This perspective is quite fitting for stochastic processes, such as those in particle physics experiments where one can repeatedly observe and count occurrences, such as particle decays, to establish probabilities. However, this frequency-based interpretation becomes problematic when applied to one-off or unique events where repetition under identical conditions

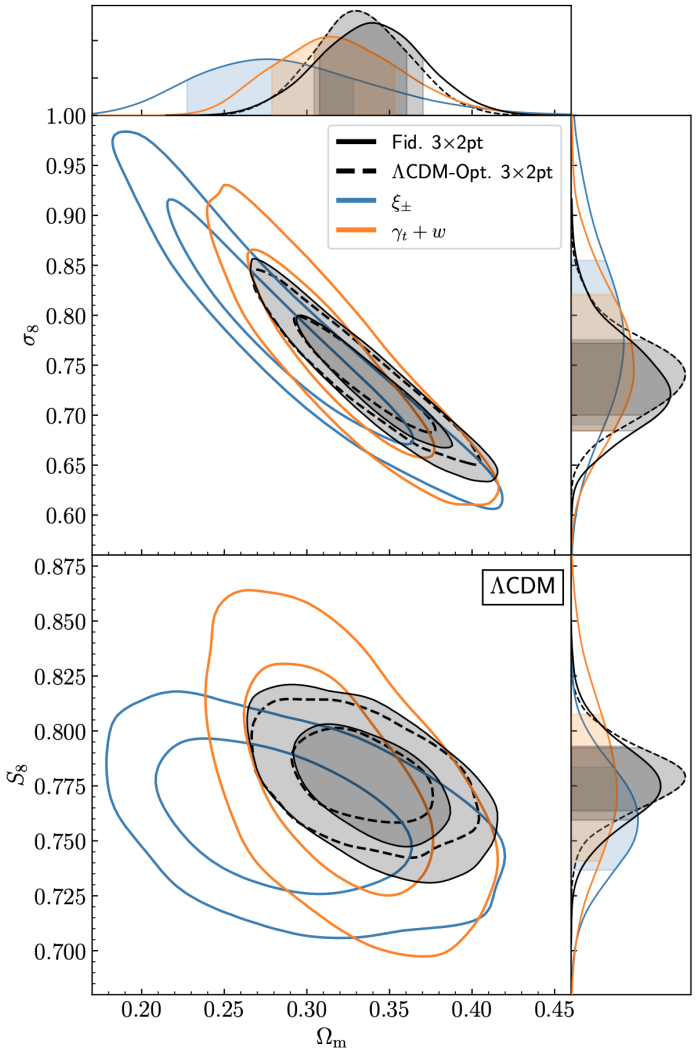

Figure 1.5: Taken from DES Collaboration (2022): Marginalized constraints on the three parameters σ_8, $S_8 = \sigma_8\sqrt{\Omega_m/0.3}$, and Ω_m in the ΛCDM model from cosmic shear ($\xi_\pm$, blue), galaxy clustering and galaxy–galaxy lensing ($\gamma_t + w(\theta)$, orange) and their combination (3×2pt, solid black). We also show a ΛCDM-optimized 3×2pt analysis that is valid for ΛCDM using smaller angular scales in cosmic shear (dashed black). The marginalized contours in this and further figures below show the 68% and 95% confidence levels. The top and side panels show 1D marginalized constraints with the 68% confidence region indicated.

is impossible or does not make much sense. For instance, determining the probability of a singular event in the future or estimating cosmological parameters, which are fixed attributes of our singular universe, does not lend itself well to a frequency-based approach. In these cases, repeating measurements would not yield different outcomes as these are not variables subject to randomness.

Bayesian statistics offers a more nuanced approach, conceptualizing probability as a measure

of an observer's belief or confidence in a particular event or hypothesis. Rather than envisioning countless iterations across multiple universes, Bayesian probability allows for a single reality but quantifies the observer's confidence level. For example, stating a probability of an event as 25% is not claiming that the event will occur in one out of four identical universes. Instead, it is expressing that, given the existing information and understanding, there is a moderate level of skepticism about the event's occurrence. This interpretation underpins much of contemporary scientific investigation, providing a framework for incorporating new evidence and updating beliefs or hypotheses accordingly.

Consider a hypothesis A and a set of background information I which encapsulates all our existing knowledge, including the universe and its physical laws. The assertion that $P(A, I) = P(I, A)$ is fundamentally true. We can then formulate this relationship in terms of the probability of A given I, $P(A|I)$, as

$$P(A|I)P(I) = P(I|A)P(A), \tag{1.60}$$

this equation is an universal expression of Bayes' theorem, and is often written as

$$P(A|I) = \frac{P(I|A)P(A)}{P(I)}. \tag{1.61}$$

Consider an observation, labeled as $\boldsymbol{D}$, which might be one of the two-point correlation functions from the previous section, like shear across angular scales. This observation, a series of scalar values arranged in a vector, is influenced by the set of cosmological parameters denoted by $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$. These parameters are chosen based on the cosmological model being tested, the capabilities of the observing instrument, and the nature of the data collected. For a given set of parameters, our model M predicts what we should expect to see in our data $\boldsymbol{D}$. Our goal is to derive the probability distribution of the parameters $\boldsymbol{\theta}$ given our data $\boldsymbol{D}$ and the model M. This is achieved through Bayes' theorem:

$$P(\boldsymbol{\theta}|\boldsymbol{D}, M) = \frac{\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M)}{P(\boldsymbol{D}|M)}. \tag{1.62}$$

In this formula, $P(\boldsymbol{\theta}|M)$ represents the prior probability, which constrains $\boldsymbol{\theta}$ based on prior knowledge or physical principles. The denominator, $P(\boldsymbol{D}|M)$, also known as the evidence, is constant over the parameter space and hence does not influence the posterior distribution $P(\boldsymbol{\theta}|\boldsymbol{D}, M)$, although it is crucial for comparing different models. The likelihood, $\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M)$, assesses the chance of observing the data $\boldsymbol{D}$ given the parameters $\boldsymbol{\theta}$. This likelihood is typically calculated by comparing the observed data $\boldsymbol{D}$ to the model predictions $\boldsymbol{T}$, usually under the assumption of Gaussian errors, leading to:

$$\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M) = \frac{1}{\sqrt{2\pi}|\boldsymbol{C}|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{D} - \boldsymbol{T}(\boldsymbol{\theta}, M))^{\top}\boldsymbol{C}^{-1}(\boldsymbol{D} - \boldsymbol{T}(\boldsymbol{\theta}, M))\right], \tag{1.63}$$

where $\boldsymbol{C}$ denotes the covariance matrix of the data, and $|\boldsymbol{C}|$ is its determinant. Notice that the determinant of the covariance matrix in denominator of the likelihood function plays a crucial role in normalizing the probability distribution. However, for the purpose of parameter estimation,

this determinant can often be ignored if it is independent of the parameters being estimated, as it then acts as a constant and does not influence the maximum of the likelihood, which determines the best-fit parameters. Cosmological analyses currently assume this independence and validate it by recalculating the covariance at the best-fit parameters, repeating the analysis, and checking for consistency in the resulting parameters. Therefore, simplifying, we get:

$$-2 \ln \mathcal{L} = 2\chi^2 = \boldsymbol{R}^\top \boldsymbol{C}^{-1} \boldsymbol{R}, \tag{1.64}$$

with $\boldsymbol{R}$ defined as the residual between the data and model predictions, $\boldsymbol{R} \equiv \boldsymbol{D} - \boldsymbol{T}$. The computation and inversion of $\boldsymbol{C}$, especially when it is large, can be computationally intensive. With the covariance matrix in hand, estimating parameters boils down to iteratively calculating these equations for different sets of $\boldsymbol{\theta}$.

For models with a small number of parameters, $P(\boldsymbol{\theta}|\boldsymbol{D}, M)$ can be computed across a grid in the parameter space. However, for models with many parameters, methods such as Markov Chain Monte Carlo (MCMC), Importance Sampling, and adaptive techniques like MultiNest (Feroz et al., 2009, 2019) or PolyChord (Handley et al., 2015a,b) are more computationally feasible and effective.

In cosmology, it is often more useful to ascertain a confidence region – a region of plausible parameter values – than to rely on a singular maximum likelihood estimate. Often, these parameters are a mix of the cosmological parameters we are interested in, and those representing systematic errors. To refine our focus on the essential parameters, we employ *marginalization*. Consider a study examining two parameters: the primary interest a_1 and the secondary a_2, possibly related to systematic bias. If they are independent, a_2 can be disregarded, but if they are degenerate and intertwined, we must consider their joint probability distribution,$P(a_1, a_2)$. To remove the effect of a_2, we integrate over all its possible values against each value of a_1

$$P(a_1) = \int P(a_1, a_2) \mathrm{d}a_2. \tag{1.65}$$

This effectively simplifies our analysis to a one-dimensional probability distribution for a_1 alone, thus marginalizing over a_2.

Goodness of fit and χ^2 metric

The concept of *goodness of fit* and the *chi-square* (χ^2) test are essential tools in statistics for assessing how well a theoretical distribution fits an observed set of data. This test is particularly designed to determine whether there is a significant match between the observed distribution of data and what is expected theoretically.

We defined the χ^2 in Equation 1.64 as

$$\chi^2 = \frac{1}{2} \boldsymbol{R}^\top \boldsymbol{C}^{-1} \boldsymbol{R}, \tag{1.66}$$

where it is clear its relation to the residuals $\boldsymbol{R}$ – the discrepancies between what we observe and what our model predicts – and the covariance matrix $\boldsymbol{C}$, which accounts for variances and

covariances among data points. The χ^2 statistic is a scalar value, indicating the fit of the model to the observed data. This value is then divided by the effective number of degrees of freedom, which depends on the number of parameters that a data set ends up constraining compared to the priors it began with.

If the χ^2 statistic is too high, it suggests that the model may not be an adequate representation of the data. If it's too low, it might indicate that the model is over fitting the data or that the errors have been overestimated. It is important to note that the χ^2 test assumes that the errors are normally distributed and that the model is correct. If these assumptions are not met, the χ^2 test may not be valid.

1.3.2 Machine Learning

Machine learning has found significant application in cosmology. This subsection delineates an array of machine learning methodologies, for a more detailed discussion on these methods, readers are referred to comprehensive texts such as Bishop (2006), Goodfellow et al. (2016). Machine learning paradigms are categorized into supervised learning, unsupervised learning, and reinforcement learning.

In supervised learning, algorithms process labeled data sets for predictive modeling and classification. Techniques such as decision trees, ensemble methods (Random Forests and Gradient Boosting Machines), Support Vector Machines, and Neural Networks are utilized for a variety of problems, usually under the assumption that some "truth" sample is available for comparison with the results of the regression or classification method. Conversely, unsupervised learning algorithms are applied to unlabeled data sets to infer patterns and gain new insights. Techniques such as Principal Component Analysis, Hierarchical Clustering and Self-Organizing Maps are employed for dimensionality reduction and clustering. In these cases, the "truth" is not available, and the algorithm itself has to learn the information contained in the data. Reinforcement learning, characterized by algorithms optimizing decision-making through environmental interactions, holds prospective applications in cosmology, such as the optimization of telescopic survey strategies.

Within this thesis, some machine learning techniques are leveraged, with a particular focus on self-organizing maps (SOMs) for the analysis and interpretation of photometric data. SOM is a type of neural network that uses the unsupervised learning approach. Its architecture is simpler than that of a deep neural network, composed of multiple layers, since it essentially has only the input and output layers. The SOM effectively reduces high-dimensional data into two-dimensional representations, while preserving the original topology through the concept of neighborhood between its nodes. Figure 1.6 shows an schematic representation of a SOM.

The SOM employs an iterative learning process where network nodes adjust their weights to align more closely with input vectors, influenced by both the node's weight vector and its distance from the best matching unit. Through iterative training, the network forms a map with adjacent nodes representing similar data vectors. SOMs provide an efficient mechanism for the reduction and interpretation of the extensive and high-dimensional data sets encountered in cosmology. In particular, SOMs have been applied to the problem of photometric redshifts, since it can learn

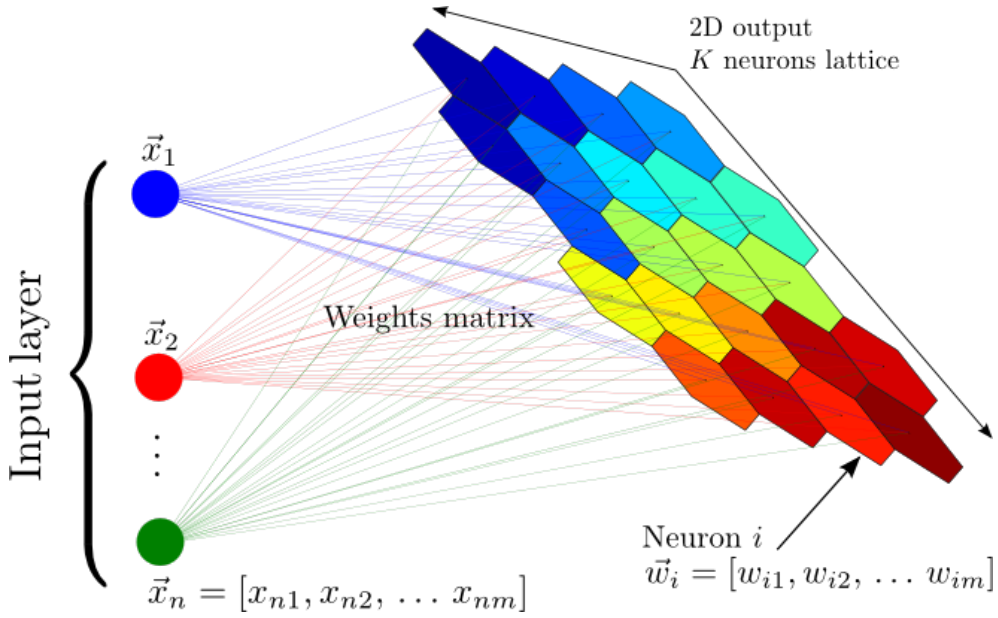

Figure 1.6: From Carrasco Kind & Brunner (2014): A schematic representation of a SOM. The training set of n galaxies, with m features each, is mapped into a two-dimensional lattice of k neurons (or cells). The weights matrix has dimensions $k \times m$, and makes the connection between the input vectors and the output map, such that each cell is associated with a weight vector. In the training phase, all neurons compete to best match each galaxy, but ultimately, each galaxy is mapped to only one neuron that most closely represents its features in the m-dimensional space. The colour of the map encodes the organization of groups of galaxies with similar properties. The main characteristic of the SOM is that it produces a non-linear mapping from an m-dimensional space of attributes (e.g. magnitudes) to a two-dimensional lattice of neurons.

the complex relations in flux (or color/magnitude) space, and effectively map it to redshift space. That makes it a power tool for redshift characterization for galaxy samples in weak lensing and galaxy clustering. We will explore SOMs in detail in Chapter 2.

1.4 Thesis Outline

This thesis presents the findings of three distinct research projects. These projects are dedicated to the development and enhancement of statistical and machine learning methodologies for the analysis of extensive data sets in large-scale structure surveys, with special focus on the problem of systematic uncertainties. While the the first two project primarily apply these methods to address weak lensing phenomena, they present enough generality to be applied for galaxy clustering, 3×2pt and a wider range of problems in large-scale structure data analysis. Each project showcases novel approaches and techniques, contributing to a more sophisticated understanding of complex cosmological data.

Chapter 2 focuses on improving one of the state of the art methods for photometric redshift characterization, a crucial component of weak lensing analysis. I investigate three modifications to the existing Self-Organizing Map (SOM) methodology used in the Dark Energy Survey Year 3 analysis. The first modification involves using a new SOM algorithm, which has a distance metric designed for the problem of photometric redshifts. The second modification involves incorporating additional information about the g-band flux, which in general has low signal-to-noise ratio (SNR) but can still provide crucial information about redshifts. The third modification involves trying to incorporate the redshifts available for some galaxies in the sample as an additional feature. I apply these modifications to the DES Year 3 data and compare the results to the original methodology, finding that the new SOM algorithm performs better, especially when combined with the addition of the g-band fluxes. This study is pivotal for the DES Year 6 analysis, since my findings lead to the adoption of the pipeline that I developed for the characterization of redshifts for the source and lens galaxies. In addition, this method can be applied to future surveys, like LSST. In fact, my pipeline is being implemented in RAIL and will be available for redshift characterization. This work is currently internal collaboration review in DES and will soon be submitted for publication as "Enhancing weak lensing redshift distribution characterization by optimizing the Dark Energy Survey Self-Organizing Map Photo-z method".

Chapter 3 addresses the problem of model selection in cosmology, which is the challenge of choosing the best model to fit observational data while avoiding biases and balancing model complexity. We propose an empirical approach that uses synthetic data to calibrate the relation between parameter bias and model complexity, and applies this method to the problem of intrinsic alignments in weak lensing surveys. We assess the bias that could be introduced due to model mispecification, in the context of the DES Y3, between the NLA and TATT models for intrinsic alignment. We also show that the level of conservatism when choosing a model can be controlled through analysis choices, which makes the method more explicit and quantifiable than other approaches to model selection. This approach can be applied to any type of data and/or systematics, and can help to choose the best model for their data, especially in the absence of informative priors. This work resulted in the publication of the paper "An empirical approach to model selection: weak lensing and intrinsic alignments", published in the "Monthly Notices of the Royal Astronomical Society, Volume 525, Pages 1885–1901"(Campos, Samuroff & Mandelbaum 2023).

In chapter 4 we address the problem of quantifying if measurements of different surveys, in this case DES and Plack, are in *tension*. I create a suite of simulated DES data sets with a controlled level of induced tension relative to the best-fitting *Planck* 2018 cosmology, and we apply a number of methods to quantify this synthetic tension and assess their performance. We also apply the same tension metrics to quantify any tension between the published constraints from the first year of DES data and the *Planck* 2015 and 2018 data sets. We recommend a strategy to evaluate if the predictions from different surveys are in accordance. We also conclude that there is evidence suggesting some tension between DES and *Planck* data, which could indicate unaccounted-for systematic effects in one or both experiments or that the underlying model is inadequate to explain the data. The results of this study have been used to inform concordance in several analyses in cosmology by providing a framework for evaluating tensions between different data sets. This work resulted in the publication of the paper "Assessing tension metrics with dark energy survey and Planck data", published in the "Monthly Notices of the Royal Astronomical Society, Volume 505, Pages 6179-6194"(Lemos, Raveri, Campos et al. 2021).

In Chapter 5, I discuss the implications of the work described in this thesis.

Chapter 2

Enhancing weak lensing redshift distribution characterization by optimizing the Dark Energy Survey Self-Organizing Map Photo-z method

Abstract

Characterization of the redshift distribution of ensembles of galaxies is pivotal for cosmological studies. In this work, we focus on improving the Self-Organizing Map (SOM) methodology for photometric redshift estimation (SOMPZ), specifically in anticipation of the Dark Energy Survey Year 6 (DES Y6) data. This data set, featuring deeper and fainter galaxies, demands adapted techniques to ensure accurate recovery of the underlying redshift distribution. We investigate three strategies for enhancing the existing SOM-based approach used in DES Year 3: 1) Replacing the Y3 SOM algorithm with one tailored for redshift estimation challenges. 2) Incorporating g-band flux information to refine redshift estimates. 3) Augmenting redshift data for galaxies where available. These methods are applied to DES Y3 data, and results are compared to the Y3 fiducial ones. Our analysis indicates significant improvements with the first two strategies, notably reducing the overlap between redshift bins. By combining strategies 1 and 2, we have successfully managed to reduce redshift bin overlap in DES Y3 by up to 66%. Conversely, the third strategy, involving the addition of redshift data for selected galaxies as an additional feature in the method, yields inferior results and is abandoned. Our findings contribute to the advancement of weak lensing redshift characterization and lay the groundwork for better redshift characterization in DES Year 6 and future stage IV surveys, like the Rubin Observatory.

gravitational lensing: weak – methods: statistical – techniques: photometric – galaxies: distances and redshifts – cosmology: observations.

2.1 Introduction

Large galaxy surveys afford us promising opportunities to learn about the constituents of the universe and the way they are distributed in space. This in turn can help us connect fundamental physics – for example of dark energy and dark matter – to observations and to learn about the nature of the most mysterious substances postulated to exist. Photometric surveys can capture images of many more galaxies than spectroscopic surveys but are hindered by the inability to measure accurate distances to the objects they image. *Photometric redshifts*, or distances inferred from the observed galaxy properties such as colors, have become essential in extracting information about cosmology from these large surveys.

One of the observables for which photometric redshifts play a major role in is weak gravitational lensing. Weak gravitational lensing is a fundamental cosmological probe that enables the investigation of the large-scale structure of the universe and has been employed in many contemporary analyses (see, e.g., Heymans et al. 2013; DES Collaboration 2016a; Jee et al. 2016; Hildebrandt et al. 2017; Troxel et al. 2018; Hikage et al. 2019; Hamana et al. 2020; Asgari et al. 2021; Loureiro et al. 2022; Amon et al. 2022; Secco, Samuroff et al. 2022; Doux et al. 2022; Dalal et al. 2023; Li et al. 2023).

In photometric surveys, while galaxy positions serve as tracers of matter density, it is by measuring the distortions in the shapes and orientations of background galaxies induced by the gravitational influence of intervening mass distributions that we can obtain a direct connection to the underlying density field. However, to extract precise cosmological information from weak lensing, it is imperative to have a robust characterization of the redshift distribution, $n(z)$, of the observed galaxies. Measuring the spectrum of each galaxy in a large optical imaging survey, though, is unfeasible, and therefore spectra are available only for small subsets of galaxies. As a result, photometric surveys heavily rely on limited, noisy photometric bands to estimate redshifts. The main challenge arises from degeneracies in the color-redshift relation, which prevent the unique determination of redshifts from wide-band photometry. The accurate characterization of the redshift distribution thus becomes one of the main challenges, and yet a crucial aspect, for interpreting gravitational lensing measurements, including cosmic shear and galaxy-galaxy lensing correlation functions.

Techniques to estimate photometric redshifts date back several decades. Template-fitting methods compare the observed photometric data of galaxies with a library of template spectra, allowing for redshift estimation (Benítez, 2000; Ilbert et al., 2009). However, this approach can be sensitive to template choices and might not capture all spectral features accurately, leading to biases in redshift predictions, particularly for poorly represented galaxy populations. Empirical approaches exploit statistical correlations between observable features (e.g., color-redshift relations) to estimate photometric redshifts (Blake & Bridle, 2005; Mandelbaum et al., 2008). However, these methods necessitate accurate and extensive spectroscopic data for calibration. Machine learning techniques, such as artificial neural networks or random forests, have gained popularity due to their ability to learn complex photometry-redshift relationships from training data sets (Collister & Lahav, 2004; Carrasco Kind & Brunner, 2013). Nonetheless, these methods heavily rely on the quality and representativeness of the training data, and their performance can

degrade when extrapolating to redshift regimes not adequately covered by the training set. Most recently, unsupervised machine learning methods that compress data embedded in a Bayesian approach have emerged as a promising direction (see for example, Buchs et al., 2019).

A Self-Organizing Map (SOM), also known as a Kohonen map (Kohonen, 1982), is an unsupervised machine learning algorithm and neural network architecture used for dimensionality reduction and data mining. It allows for complex and high-dimensional data to be represented in a lower-dimensional space while preserving the topological relationships between data points. For the purposes of redshift estimation, when assigning each galaxy to a cell in a Self Organizing Map (SOM), galaxies with similar redshifts are grouped in the same cell, or "nearby" cells if the grouping is in a 2D grid, and the redshift distribution for those galaxies can be determined fairly accurately. DES used this technique in its Year 3 weak lensing cosmological analyses (Myles, Alarcon et al. (2021); DES Collaboration (2022); Amon et al. (2022); Secco, Samuroff et al. (2022)) and KiDS has used it (Wright et al., 2020) to achieve few-percent level constraints on the mean of the redshift distribution for each redshift bin. It has emerged as a viable candidate for upcoming surveys such as Rubin and Euclid (Ivezić et al., 2019; Laureijs et al., 2011), but improvements are required to achieve sub-percent level constraints (The LSST Dark Energy Science Collaboration et al., 2018; Euclid Collaboration et al., 2020).

Here, we explore several improvements to the SOM methodology used in DES-Y3, ahead of the final DES Year-6 (Y6) analysis. This serves two primary purposes: (i) allowing for the potential of improving on the Y3 implementation and (ii) stress-testing the robustness of the cosmological conclusions. The latter point is particularly important in the context of more stringent requirements that come with more statistically powerful data, as well as applying this methodology to deeper photometric data. If different implementations of the SOM framework give the same answer, we will become more confident applying it moving forward as statistical errors continue to decrease.

First, we test replacing the SOM algorithm used in Y3 by the one proposed in Sánchez, Raveri, Alarcon & Bernstein (2020). This new algorithm implements a Self-Organizing Map with a distance metric specific for the problem of photometric redshift estimation. Although it was shown in the Year 3 analysis that the generic SOM algorithm is already successful at estimating redshifts at the percentage level (Myles, Alarcon et al. 2021), we hope that by introducing a SOM that is tailored for the problem of redshifts, we can achieve even better precision. Second, we show the impact that including an extra flux band, the g-band, has on our ability to obtain well-defined redshift bins, motivating the importance of well calibrated point spread functions in those limits, such that we do not lose this very crucial piece of information. Finally, we try adding the redshift information of the spectroscopic galaxies in our sample as an additional feature in the SOM. This is an unconventional approach to a unsupervised machining learning method, since the norm is for the quantity being estimated to not be part of the features, and indeed we find that it is not beneficial, but we present our attempts for the sake of completeness.

Section 3.3 details the DES Year 3 data that we re-analyse with the proposed modifications to our redshift estimation method. Section 2.3 presents a summary of the Self-Organizing Map algorithm and the SOMPZ method for redshift estimation used by DES. Section 2.4 presents the proposed modifications to the SOMPZ method used in DES Y3. Section 2.5 discusses the results

Table 2.1: Summary of the catalogs used in DES Year 3 for redshift estimation of the weak lensing source galaxies, including the area covered and the number of galaxies.

Sample	Area (sq. deg.)	Number of Galaxies
Wide	4143	100,208,944
Deep	5.88	2.8M
Redshift	-	57,000
Deep/Balrog	-	2,417,437

of implementing these different modifications, and their impact on the redshift bins. Finally, Section 2.6 shows the impact on cosmological parameter constraints.

2.2 The Dark Energy Survey

We summarize the samples used in this work in Table 2.1. These are the same ones used for redshift characterization of the weak lensing source galaxies in DES Year 3. The strategy employs Self-Organizing maps (SOMs), which we detail in Section 2.3, and leverages the information present in three catalogs - wide, deep, and redshift - as well as Balrog injections:

Wide: The weak lensing source catalog is described in detail in Gatti et al. (2021). After the applied selections $\hat{s}$ in magnitudes and colors, the *wide sample* is composed of $100,208,944$ galaxies, spread over 4143 square degrees. DES has made flux measurements for all of these galaxies in the *griz* bands of the electromagnetic spectrum (although the g-band was not used in DES-Y3).

Deep: The *deep sample* refers to the DES deep field galaxies, which have measured fluxes in additional bands *ugrizJHK*. There are four deep-fields mapped in DES Y3, see Hartley et al. (2021), that added cover an area of 5.88 square degrees. Notice that Y-band data in the deep fields had large offsets between the constituent exposures, and therefore could not be used.

Redshift: A subset of the deep field galaxies have accurate redshifts (Myles, Alarcon et al. 2021) obtained from a variety of external data sets. We call this set, containing $57,000$ galaxies, the *redshift sample*.

Balrog: In order to connect the information in our samples, we use Balrog injections. The Balrog software, developed by Suchyta et al. (2016), enables the creation of simulated galaxies, or Balrog injections, which are inserted into authentic images. These synthetic galaxies are designed based on the DES deep field photometry and are placed multiple times at various positions across the broader wide-field footprint, as specified in Everett et al. (2022). The resulting catalogue, called the *deep/Balrog* sample, includes $2,417,437$

injection-realization pairs, each of which has both deep and wide photometric data. This sample is a crucial element of our redshift calibration inference technique.

In what follows, we denote the wide data by $\hat{\boldsymbol{x}}$ with covariance matrix $\hat{\Sigma}$ and the deep data by $\boldsymbol{x}$ with covariance Σ, and the selection by $\hat{s}$, following the notation in Buchs et al. (2019); Myles, Alarcon et al. (2021). The wide field data vector has three components, $\hat{\boldsymbol{x}} = [r, i, z]$. For the deep fields, there are 3 infrared bands available – J,H, and K – and the DES u, g bands are also used, such that $\boldsymbol{x} = [u, g, r, i, z, J, H, K]$ has eight components in total. Since the redshift galaxies are a subset of the deep galaxies, they too have the 8 components $\boldsymbol{x}$; Balrog galaxies typically have approximately 15 realizations $\hat{\boldsymbol{x}}$ (corresponding to the number of wide field injections) for a single $\boldsymbol{x}$ (corresponding to a single deep field galaxy).

2.3 Self-organizing maps for photometric redshifts

In what follows, we review the SOM standard algorithm, and describe the SOMPZ method, i.e., how SOMs are used in practice for redshift estimation in DES-Y3.

2.3.1 The SOM Algorithm

A Self-Organizing Map (SOM) is a type of Artificial Neural Network (ANN) that produces a discretized, lower dimensional, representation of the input space, while preserving its topology. Proposed by Kohonen (1982), it is an unsupervised Machine Learning method that uses soft competitive learning: the cells of the map (also known as nodes or neurons) compete to most closely resemble each training example until the best matching unit (BMU) is found, then the winner and its neighborhood are adapted.

Consider a set of n training samples, each with m features, i.e., for each sample we have an input vector $\boldsymbol{x} \in \mathrm{R}^m$. In our case, for instance, each sample is a single galaxy and the features are its fluxes (or colors or magnitudes) in m bands. The SOM can be understood as collection of C cells arranged in a l-dimensional grid that has a specified topology. Each cell is associated with a weight vector $\boldsymbol{\omega}_k \in \mathrm{R}^m$, where $k = 1, \cdots, C$. Both the input and weight vectors live in the input space, while the cells live in the output, or lattice, space.

The training of a SOM is relatively simple. The weights are initialized to random or from data samples and the learning happens in three stages: Competition, cooperation, and weight adaptation.

- Competition: at each step, a random sample of the training set is presented to the self-organizing map. The cell whose weight vector is the closest to the sample vector is identified as the best matching unit (BMU):

$$c_b = \underset{k}{\operatorname{argmin}}\, d(\boldsymbol{x}, \boldsymbol{\omega}_k). \tag{2.1}$$

The "closeness" is measured by some distance metric $d(\boldsymbol{x}, \boldsymbol{\omega}_k)$ between the sample and the cell in the map. Typically the Euclidian distance is used, but in Buchs et al. (2019) the chi-square distance was chosen:

$$d^2(\boldsymbol{x}, \boldsymbol{\omega}_k) = (\boldsymbol{x} - \boldsymbol{\omega}_k)^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\omega}_k), \tag{2.2}$$

where $\boldsymbol{\Sigma}$ is the covariance matrix for the training vector, $\boldsymbol{x}$. The cell minimizing this distance is identified as the BMU, and the sample is then assigned to it.

- Cooperation: to preserve the topology of the input space, not only the BMU is identified and updated, but also its neighborhood. The neighborhood function $H_{b,k}(t)$ creates the connection between the input space and the cells in the map, responsible for the self-organizing property of the map. The size of the BMU's neighborhood decreases as a function of time steps, t. In addition, $H_{b,k}$ should decrease as the distance from the BMU increases. It must also satisfy the properties that it is maximum in the winning cell b and is symmetric about it. A Gaussian neighborhood function attend those requirements:

$$H_{b,k}(t) = \exp[-D^2_{b,k}/\sigma^2(t)]. \tag{2.3}$$

The distance between the BMU, c_b, and any cell on the map, c_k, is the Euclidian distance in the l-dimensional map:

$$D^2_{b,k} = \sum_{i=1}^{l} (c_{b,i} - c_{k,i})^2. \tag{2.4}$$

The width of the Gaussian kernel is given by

$$\sigma(t) = \sigma_s^{1-t/t_{max}}. \tag{2.5}$$

At the beginning of the training, σ_s should be large enough that most of the map is initially affected. As the training progresses, the width shrinks until only the BMU and its closest neighbours are significantly affected by new data.

- Weight adaptation: once the BMU is computed, we can calculate the updated value of the weight vectors for the $t + 1$-th iteration through the following relation:

$$\boldsymbol{\omega}_k(t+1) = \boldsymbol{\omega_k}(t) + a(t) H_{b,k}(t) [\boldsymbol{x}(t) - \boldsymbol{\omega}_k(t)], \tag{2.6}$$

where t is the current time step in training, $a(t)$ is the learning rate:

$$a(t) = a_0^{t/t_{max}}, \tag{2.7}$$

where $a_0 \in [0, 1]$. In each iteration, this update function is applied to each of the cells in the map.

These steps describe the standard SOM algorithm, which has been applied for the purpose of redshift estimation in previous works (see e.g. Masters et al. (2015); Speagle et al. (2019); Buchs et al. (2019)), including the DES-Y3 analyses (Myles, Alarcon et al. 2021).

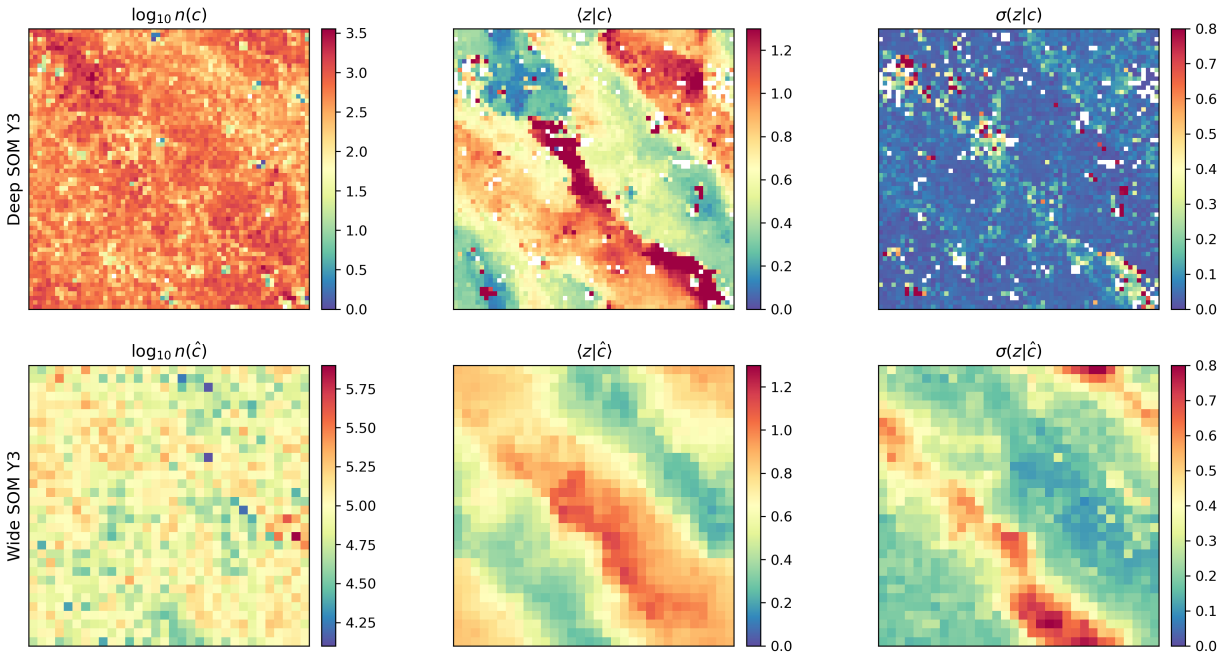

Figure 2.1: Visualization of the self-organizing maps constructed using the fiducial Y3 SOM algorithm described in Section 2.3.1. Top: Deep field self-organizing map composed of 4096 cells. Bottom: Wide field self-organizing map composed of 1024 cells. The left-hand panels show the total number of galaxies assigned to each SOM, the middle panels show the mean redshift for each cell, and the right panels show the standard deviation of the redshift distribution in each cell of the map. The white cells found in the deep SOM are due to the lack of spectroscopic information in those regions of the color space, i.e., there are no galaxies in the COSMOS2015 sample that were assigned to those cells.

2.3.2 Dark Energy Survey SOMPZ

In order to estimate the redshift distribution of the wide sample, we construct two SOMs:

1. Deep SOM: trained using the deep data, to which we assign the deep galaxies (and their subset of galaxies with redshifts);

2. Wide SOM: trained using a random sub-sample of the wide data, to which we assign the wide sample, and the Balrog injections of the deep samples.

Figure 2.1 shows the number of objects, the redshift distribution and the standard deviation in the deep (top) and wide (bottom) SOMs used in the DES Y3 analysis. As we can see, and the name suggests, the deep SOM has deeper, and fainter, galaxies, going to higher z values compared to the wide SOM.

Each SOM cell acts as a way of discretizing the continuous colour and colour-magnitude spaces spanned by $\boldsymbol{x}$ and Σ (and $\hat{\boldsymbol{x}}$, $\hat{\Sigma}$) into discrete categories c and $\hat{c}$. Therefore, the probability

distribution function for the redshift of an ensemble of galaxies, conditioned on being observed in a particular cell $\hat{c}$, and on passing a selection function $\hat{s}$, can be written by marginalizing over the deep-field information:

$$p(z|\hat{c},\hat{s}) = \sum_{c} p(z|c,\hat{c},\hat{s})p(c|\hat{c},\hat{s}). \tag{2.8}$$

We then assign each cell $\hat{c}$ to a tomographic bin (see Myles, Alarcon et al. 2021 for the details on the assignment algorithm) and construct the $n(z)$ of each bin by summing over the cells belonging to the bin:

$$n_{\hat{b}}(z) \equiv p(z|\hat{b},\hat{s}) = \sum_{\hat{c}\in\hat{b}} p(z|\hat{c},\hat{s})p(\hat{c}|\hat{s},\hat{b}) \tag{2.9}$$

$$= \sum_{\hat{c}\in\hat{b}}\sum_{c} p(z|c,\hat{c},\hat{s})p(c|\hat{c},\hat{s})p(\hat{c}|\hat{s},\hat{b}). \tag{2.10}$$

In the equation above, each term is obtained from one of the galaxy samples we are using:

1. $p(z|c,\hat{c},\hat{s})$ is computed from the redshift sample subset of the deep sample, which contains spectroscopic redshifts, deep photometry, and wide-field Balrog realisations. It tells us the probability of getting a redshift z, given the deep cell c, the wide cell $\hat{c}$, and the selection $\hat{s}$.

2. $p(c|\hat{c},\hat{s})$ is computed from the Balrog injections of the entire deep sample. It tells us the probability of ending up in the deep cell c, given the wide cell $\hat{c}$ and the selection $\hat{s}$. We call this term the transfer function, because it connects the deep and wide cells. It is computed from Balrog realisations, because it requires both wide-field and deep-field photometry to be available.

3. $p(\hat{c}|\hat{s},\hat{b})$ is computed from the wide sample. It tells us the probability that a galaxy in bin $\hat{b}$ is in the wide SOM cell $\hat{c}$. Therefore, cells with very few galaxies in them are down-weighted when determining the redshift distribution of the bin.

Assuming that the $p(z)$ in the deep cells (with high quality photometry) do not depend on the wide (noisy) photometry of those galaxies, we can remove the conditions on $\hat{c}$ and $\hat{b}$ in the first and last terms of Equation 2.10, and approximate it to

$$p(z|\hat{b},\hat{s}) \approx \sum_{\hat{c}\in\hat{b}}\sum_{c} p(z|c,\hat{s})p(c|\hat{c},\hat{s})p(\hat{c}|\hat{s}). \tag{2.11}$$

The transfer function, $p(c|\hat{c},\hat{s})$, connecting the deep and wide samples, is computed from Balrog realisations, not the full wide galaxy sample. Re-writing it as

$$p(c|\hat{c},\hat{s}) = \frac{p(c,\hat{c}|\hat{s})}{p(\hat{c}|\hat{s})}, \tag{2.12}$$

and replacing it in the equation above, we can write each term highlighting the sample from which it is obtained

$$p(z|\hat{b},\hat{s}) \approx \sum_{\hat{c}\in\hat{b}} \sum_{c} \underbrace{p(z|c,\hat{s})}_{\text{Redshift}} \underbrace{p(c|\hat{s})}_{\text{Deep}} \underbrace{\frac{p(c,\hat{c}|\hat{s})}{p(c|\hat{s})p(\hat{c}|\hat{s})}}_{\text{Balrog}} \underbrace{p(\hat{c}|\hat{s})}_{\text{Wide}} . \tag{2.13}$$

We would like to emphasize that solving Equation 2.13 is not the final result for the Y3 $n(z)$'s, as two other pieces of information were added in: clustering redshifts and shear ratios (see Myles, Alarcon et al. 2021). However, this is the main result to which we are interested in comparing in this work. In what follows we will compare this *fiducial Y3* $n(z)$ to the one obtained by each SOM modification proposed in this paper.

2.4 Testing improved SOM methodology

In this section we describe the three modifications to the standard method, and assess the impact on the DES Y3 redshift distributions: replacing the SOM algorithm used in Year 3 (see Buchs et al. 2019) by the one proposed in Sánchez, Raveri, Alarcon & Bernstein (2020); including an extra band (g-band), even though it has low SNR; including redshifts, when available, as an additional feature to train and assign galaxies to the SOM.

2.4.1 SOM for faint galaxies - SOMF

A characteristic of the majority of machine learning methods, self-organizing maps included, is the assumption that the training data is ideal, i.e., does not contain errors. This assumption is not true in general, especially when working with empirical data. This point is addressed for the case of SOMs in Sánchez, Raveri, Alarcon & Bernstein (2020), where the authors propose a modification of the standard SOM algorithm that accounts for measurement uncertainties in the training set, with the problem of faint galaxies in mind. The basic idea is to take the errors into account such that, examples with larger measurement uncertainties will result in less change to the weights than examples with smaller uncertainties. The main modifications to the standard algorithm consist in redefining the distance measure between a training sample and a cell on the map, and the training shift through which the weights are updated.

In addition, the sample features ($\boldsymbol{x}$) and cell weights ($\boldsymbol{\omega}_k$) are converted into units of signal-to-noise ratio (SNR), specifying a maximum for the sample SNR as a means of softening the

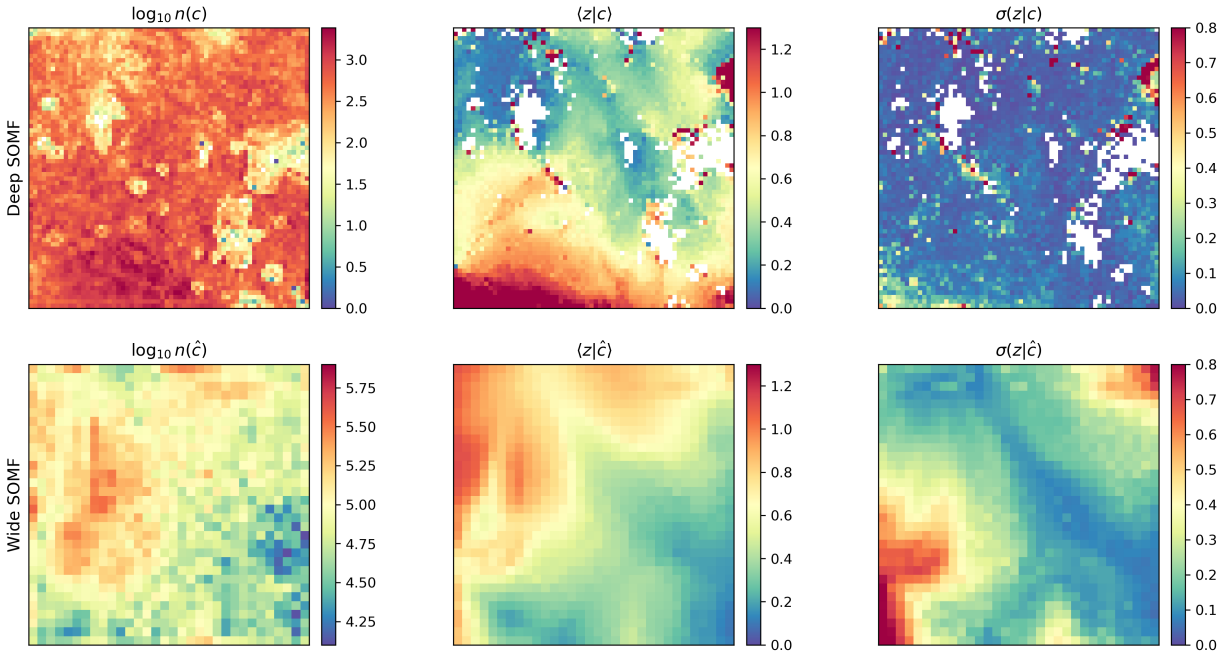

Figure 2.2: Visualization of the self-organizing maps constructed using the SOMF algorithm described in Section 2.4.1. Top: Deep field self-organizing map composed of 4096 cells. Bottom: Wide field self-organizing map composed of 1024 cells. The left-hand panels show the total number of galaxies assigned to each SOM, the middle panels show the mean redshift for each cell, and the right panels show the standard deviation of the redshift distribution in each cell of the map. The white cells found in the deep SOM are due to the lack of spectroscopic information in those regions of the color space, i.e., there are no galaxies in the COSMOS2015 sample that were assigned to those cells.

specificity of the cells:

$$s_{ib} \equiv \max\left(\Sigma_{ib}, \frac{x_{ib}}{\mathrm{SNR}_{\max}}\right) \tag{2.14}$$

$$\nu_{ib} \equiv \frac{x_{ib}}{s_{ib}} \tag{2.15}$$

$$\nu_{cb} \equiv \frac{\omega_{cb}}{s_{ib}} \tag{2.16}$$

Here, the quantities ν_{ib} and ν_{cb} are the galaxy fluxes and cell weights (indexed by the photometric band b).

The SOM algorithm presented in Section 2.3 uses the chi-square distance, defined in Equation 2.2, as the metric between the sample and the cell in the map. In Sánchez, Raveri, Alarcon & Bernstein (2020), however, the authors define:

$$d(\boldsymbol{x}, \boldsymbol{\Sigma}, \boldsymbol{\omega}_k) = \inf_s \left[\tilde{d}(\boldsymbol{x}, \boldsymbol{\Sigma}, e^s \boldsymbol{\omega}_k) + \frac{s^2}{\sigma_s^2}\right], \tag{2.17}$$

where

$$\tilde{d}(\boldsymbol{x}, \boldsymbol{\Sigma}, \boldsymbol{\omega}_k) = \sum_b \left[\frac{\operatorname{asinh} \nu_{cb} + W_{ib} \log 2\nu_{cb}}{1 + W_{ib}} - \operatorname{asinh} \nu_{ib}\right]^2 \times (1 + \nu_{ib}^2), \tag{2.18}$$

approaches the Euclidean distance in log-flux at high SNR, and is also Euclidean in linear flux at low SNR, while weighting each band by its SNR (up to a maximum). As a result this metric is better suited to the wide dynamic range of galaxy fluxes. The weighting function is defined as

$$W_{ib} = e^{2(\nu_{ib} - 4)}, \tag{2.19}$$

such that it possible to transition from the high- to low-SNR regimes. Equation 2.17 includes an overall scale constant e^s which allows the cells to be "fuzzy" in overall flux level. As pointed out in Sánchez, Raveri, Alarcon & Bernstein (2020), there is no natural periodicity in the feature space of galaxy colors and magnitudes. Therefore, the assumption of periodic boundary conditions, usual to the standard algorithm, is not adopted here.

Application to DES Y3

We test the impact of using this modified SOM methodology with the DES Y3 data. Figure 2.2 shows a deep (top) and a wide (bottom) SOM constructed using the DES Y3 data described in Section 3.3 and this modified SOM algorithm. The left panels show the number of objects distributed in each of the two SOMs. The smooth behavior of redshift across the SOM, as seen in the middle panels, shows that the variation of redshift in both the 8-band and 3-band space topology, shown in Figures 2.A.1 and 2.A.2, are reasonably well traced by the 2D SOM. The right-most panels show the standard deviation of the redshift distribution in each SOM cell.

We emphasize that, even though only fluxes are used as features, the smoothness present in the middle panel of the SOMs is evidence that the redshift space is being well mapped by the flux space. We see that the transitions between low- and high-redshifts do not happen abruptly, in general, as we expect in a successful compression of this high dimensional space. The white cells in the top panels represents cells that ended up without galaxies from the redshift sample and, therefore we could not estimate the redshift distribution in those cells. Comparing Figure 2.1 and Figure 2.2 we see that, even though the fiducial Y3 Deep SOM also had white cells, there were fewer than using the SOMF algorithm. This further emphasizes that the two SOM algorithms group galaxies from the same catalog in a different way. In a recent paper, Sánchez et al. (2022a) argue that the reason for this could be related to the SOMF algorithm being better at anomaly detection, and this difference could come from strange, undetected objects in our catalog. Notice that the map initialization when running SOM and SOMF is not the same, i.e., it is not possible to do a cell-to-cell comparison between the SOMs in Figure 2.1 and Figure 2.2.

Figure 2.3 compares the standard deviation of the redshift distribution in each cell of the wide SOM $\sigma(z|\hat{c})$ for the SOMF and the Y3 SOM, showing an overall improvement when using the SOMF algorithm. The horizontal lines represent the 25 (solid), 50 (dashed) and 75 (dotted) percentiles of $\sigma(z|\hat{c})$. We can observe that for the SOMF we have about 50% of the wide cells with $\sigma(z|\hat{c}) < 0.2$, while that is true for only about 25% of the SOM Y3 wide cells.

2.4.2 Regaining blue bands for redshift estimation - *griz*

Although measured in wide field photometry, the g-band did not have an accurate enough point spread function to measure the shapes of galaxies. In particular, the g-band rho statistics (see **?** Figure 13) were considered unacceptably large, which led to the exclusion of g-band data from the Y3 weak lensing analysis. Here we perform the exercise of including g-band information, in addition to r, i, z bands, to create and assign galaxies to the wide SOM. Notice that, since the Metacal convolution and deconvolution (Gatti et al., 2021) could not be carried out, we do not have shape measurements with g-band for Y3, therefore this exercise is purely at the photometric redshift level.

Application to DES Y3

The samples used here are exactly the same as used in obtaining the fiducial Y3 weak lensing redshift measurements (see Section 3.3), the only difference is the inclusion of the g-band in training the wide SOM and, therefore, including g-band fluxes when assigning the wide and balrog samples to the wide SOM. Our purpose is to quantify the improvement in our redshift constraints, in the hypothetical case that the g-band measurements had been considered good enough to use in the DES Y3. This is particularly timely because we expect that for the Y6 analysis the g-band PSF solution will be sufficiently improved by the addition of color dependence, allowing it to be used for the weak lensing analysis (**?**).

We test the impact of the addition of the g-band using the fiducial Y3 SOM algorithm (see Section 2.3), and the modified SOM described in Section 2.4. Adding the g-band impacts only

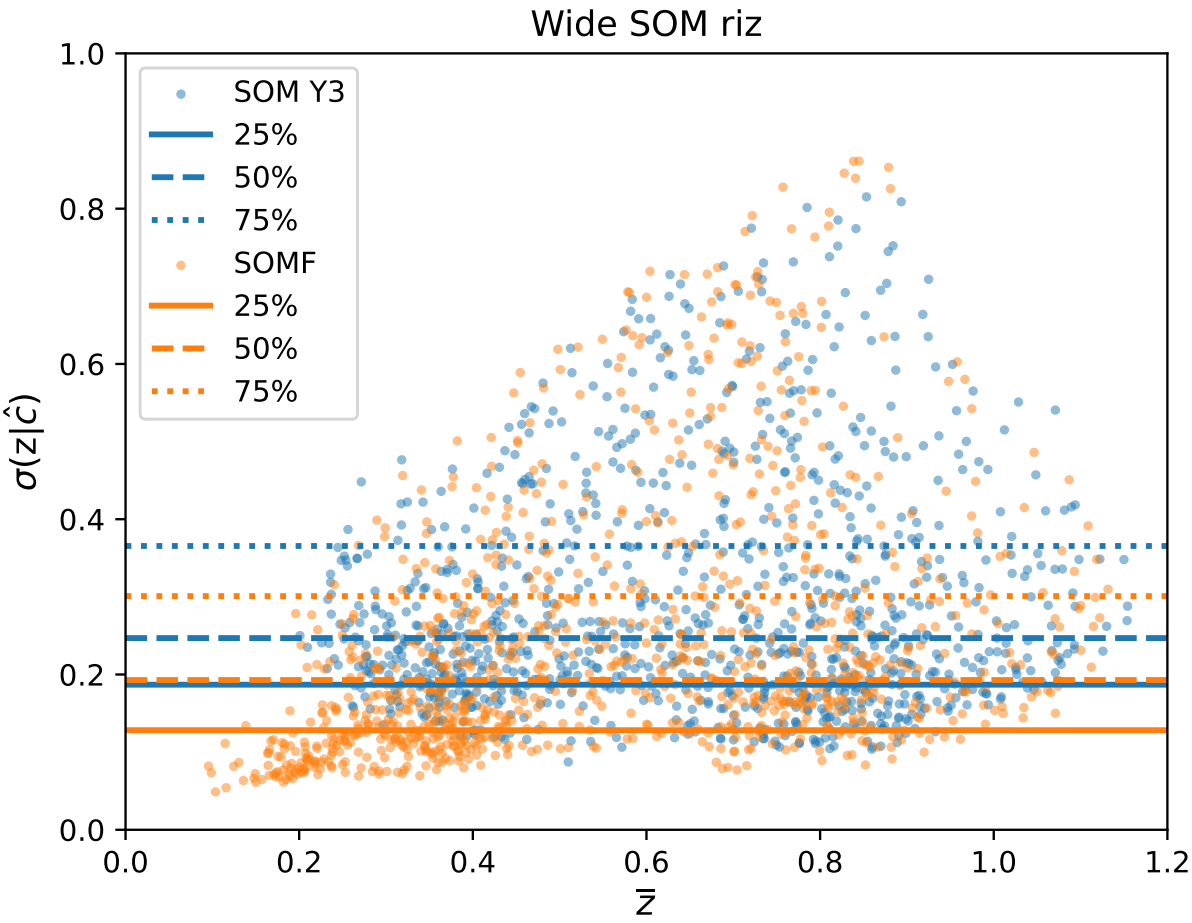

Figure 2.3: Standard deviation $\sigma(z|\hat{c})$ of the redshift distribution in each wide SOM cell, versus the mean redshift $\bar{z}$ of each cell, for the standard Y3 SOM (blue), and the SOMF (orange). The horizontal lines represent the 25 (solid), 50 (dashed) and 75 (dotted) percentiles of $\sigma(z|\hat{c})$. We can observe that the SOMF presents an overall reduction in the uncertainty per wide cell.

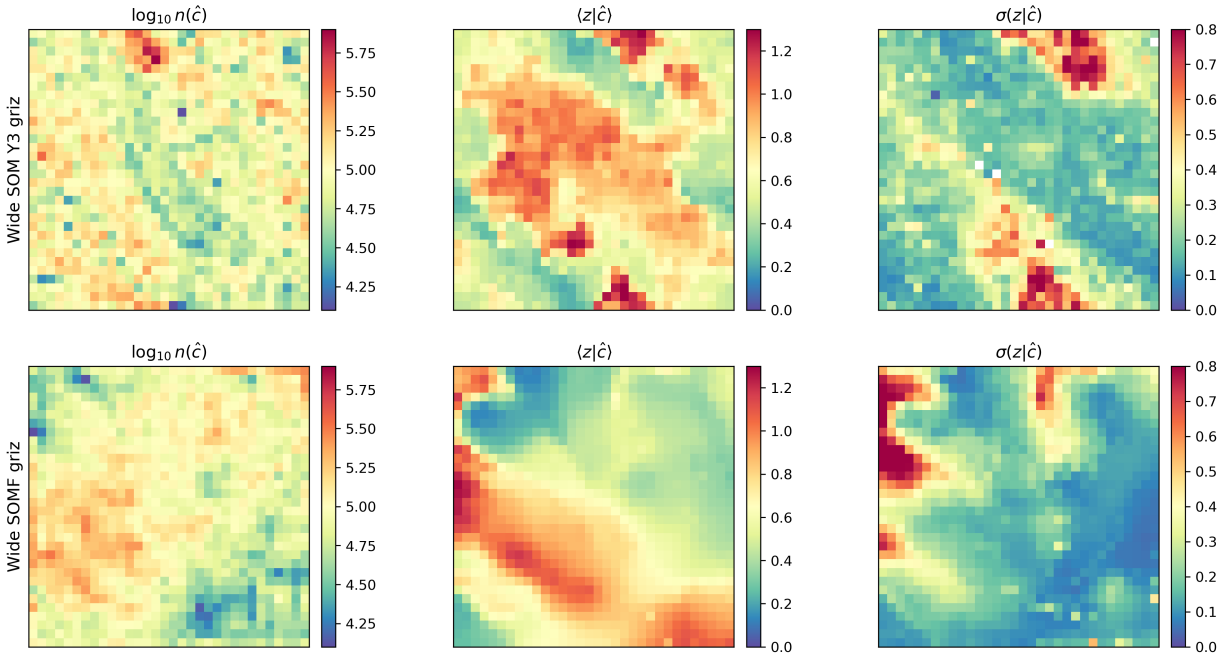

Figure 2.4: Visualization of the self-organizing maps constructed adding the g-band to train and assign the wide data, as described in Section 2.4.2. Top: Wide field self-organizing map obtained using the DES Y3 SOM algorithm, but adding the g-band information. Notice that in this case, the deep SOM is exactly the same as the Y3 one, shown in Figure 2.1. Bottom: Wide field self-organizing map obtained using the SOMF algorithm, but adding the g-band information. Notice that in this case, the deep SOM is exactly the same as the one in Figure 2.2. The left-hand panels show the total number of galaxies assigned to each SOM, the middle panels show the mean redshift for each cell, and the right panels show the standard deviation of the redshift distribution in each cell of the map.

the wide SOM part of the SOMPZ method. Figure 2.4 shows wide SOM constructed using the fiducial Y3 SOM described in Section 2.3 (top), and the one described in Section 2.4 (bottom), adding the g-band information to the train the wide SOM and assigning data to it. The left panels show the number of objects distributed in each of the two SOMs. The middle panels show the mean redshift for each cell. Notice that the inclusion of the g-band creates more cells at higher redshifts in the wide SOM, when compared to the Y3 fiducial SOM or the SOMF. The right-most panels show the standard deviation of the redshift distribution in each SOM cell, which is compatible with what we see for the Y3 fiducial SOM and the SOMF.

2.4.3 Including redshift

The original deep SOM is trained using deep field galaxies and assigned using redshift sample galaxies and deep field galaxies. This enables us to infer the deep field galaxy $p(z|c, \hat{s})$ distribution from the redshift sample galaxies for each cell. The redshift itself is *not* used as a feature; only the photometric fluxes in each band are in the feature vector. Here, we investigate the impact of including redshifts of galaxies (when available) as an extra feature, such that for each sample we have an input vector $\boldsymbol{x}$:

$$\boldsymbol{x} = [u, g, r, i, z, J, H, K, \text{redshift}] \,, \tag{2.20}$$

containing the 8 fiducial bands plus the redshift information.

To quantify the contribution of redshift in the training and assigning process, we use a weighting factor λ to modulate how much it contributes relative to the photometric bands. When $\lambda = 1$, the redshift information is normalized to have the same contribution as a luptitude. We also consider $\lambda = 0.1$ and 0.05, in which cases redshift contributes only 10% and 5% that of a luptitude, so that the redshift information plays a smaller role in constructing the SOM and assigning galaxies to cells.

The methodology and samples used here are exactly the same as used in obtaining the fiducial Y3 weak lensing redshift measurements (see Sections 3.3 and 2.3). The only difference is the inclusion of the redshift of galaxies as an extra feature in training the deep SOM. We test two possibilities (i) including redshift information both in training and assigning galaxies to the (i) both in SOM training and assigning and (ii) in the SOM training process only.

2.5 Redshift Bins and Bin Overlap

Having well-defined redshift bins is essential for weak lensing analysis, in order to ensure accurate and unbiased measurements of the gravitational lensing effect. Minimizing the overlap of redshift distributions, reduces the contamination of signals between bins, which is crucial to probe the lensing signal as a function of source redshift, control systematic errors, and enable precise cosmological parameter constraints.

Using the DES Y3 data, described in Section 3.3, we assessed the ability of the modifications to the fiducial Y3 SOMPZ method (see Section 2.3 and Myles, Alarcon et al. 2021), detailed in Section 2.4. In particular, replacing the fiducial Y3 SOM by the SOMF and including the g-band information improves our redshift constraining power and reduces the bin overlap. We detail our findings regarding those modifications in what follows. The third possibility that we described in Section 2.4.3, adding the redshift information, when available, to train and assign galaxies to the SOM, does not provide any improvements in comparison to the Y3 results. Therefore, we move our findings on that to Appendix 3.A.

	Bin 0	Bin 1	Bin 2	Bin 3
z^{PZ} range	0.0—0.358	0.358—0.631	0.631—0.872	0.872—2.0
$\langle z \rangle$ Y3 SOM	0.335	0.518	0.750	0.936
$\langle z \rangle$ SOMF	0.327	0.510	0.735	0.928
$\langle z \rangle$ Y3 SOM $griz$	0.328	0.473	0.729	0.968
$\langle z \rangle$ SOMF $griz$	0.312	0.467	0.725	0.976
Uncertainty*				
Shot Noise & Sample Variance	0.006	0.005	0.004	0.006
Redshift Sample Uncertainty	0.003	0.004	0.006	0.006
Balrog Uncertainty	< 0.001	< 0.001	< 0.001	< 0.001
Photometric Calibration Uncertainty	0.010	0.005	0.002	0.002
Inherent SOMPZ Method Uncertainty	0.003	0.003	0.003	0.003
Combined Uncertainty: SOMPZ (from 3sDir)	0.012	0.008	0.006	0.009

* We refer to Myles, Alarcon et al. (2021) for the definition of each uncertainty.

Table 2.2: Values of and approximate error contributions to the mean redshift of each tomographic bin. Given that the the only difference between the redshift distributions estimated using the Y3 SOM and the SOMF comes from the SOM recipe (all the samples are the same in both cases), we can safely assume that the uncertainties due to Shot Noise & Sample Variance, Redshift Sample, Balrog and Photometric Calibration are exactly the same ones estimated for DES Y3 (Myles, Alarcon et al. 2021). The only uncertainty affected by the change in our method is the inherent SOMPZ Method uncertainty. Figure 2.3 suggests that uncertainty to be even smaller for the SOMF, therefore we decided to not re-compute the SOMPZ uncertainty, and assume its upper bound to be the same as the Y3 SOM.

2.5.1 $N(z)$ distributions

In Figure 2.5, we compare the result of applying the redshift schema described in Section 2.3.2, in particular the solution of Equation 2.10, using both the Y3 SOM and the SOMF. We can see that the two methods agree very well, both in mean redshift and shape of the $n(z)$. In particular, once we apply the uncertainties due to each component of the method, shown in Table 2.2, all bins agree well inside the uncertainty level, with the exception of bin 2 that is slightly off the uncertainty bound. Notice that we can safely assume that the uncertainties due to shot noise and sample variance, redshift sample, Balrog and photometric calibration are exactly the same ones estimated for DES Y3 (see Myles, Alarcon et al. 2021 for details on each uncertainty and how they were estimated). The inherent SOMPZ Method uncertainty is the only one affected by the change of method, but given the good agreement in mean redshift and shape of the distributions, we decided to not recompute it, and assume it is the same as for Y3 as well. Notice that the SOMF method produces bins seem slightly better defined, with higher peaks, even though the two distributions follow each other very closely.

Figure 2.6 shows a similar comparison, but now including the g-band information, for for the fiducial Y3 SOM and the SOMF. We can see that two SOM algorithms again agree very well, both in mean redshift and shape of the $n(z)$. The mean redshift in each bin agree within the uncertainty level in Table 2.2. The difference in the peak heights, is even more pronounced now with the addition of the g-band information, showing that the SOMF algorithm leverages the g-band information to get even better defined redshift bins. Notice that the means and shapes of the distributions in Figure 2.5 and Figure 2.6 differ from each other, which is a expected consequence of the addition of the g-band information.

2.5.2 Bin overlap

We aim for minimal overlap between bins, indicating distinct redshift ranges that have been well separated. This results in a higher likelihood that a galaxy is correctly assigned to its designated bin rather than to a neighboring one. Figure 2.7 compares the amount of bin overlap obtained with each method. The amount of bin overlap when using the Y3 SOM and the *riz* bands is shown in blue, the SOMF with *riz* bands is shown in green, Y3 SOM with *griz* bands in yellow, and the SOMF with *griz* bands in red. We can immediately see that the Y3 SOM *riz* presents the highest overlap among all methods, and the greatest reduction in bin overlap is obtained when we combine the SOMF recipe and the *griz* bands.

The numerical values corresponding to the bin overlap between bin pairs in shown in Table 2.3, where we also show the percent decrease in bin overlap relative to the Y3 SOM *riz*. The SOMF *riz*, Y3 SOM *riz*, SOMF *griz* present decreasing amount in overlap, having a reduction of 3%, 23%, and 25% respectively for bins 0-1; 5%, 14%, and 33% for bins 1-2; 0%, 52%, and 66% for bins 1-3; 6%, 14%, and 31% for bins 2-3. In the case of the overlap between bins 0-2 and 0-3, the amount of overlap is already small when compared to the other bin pairs. For those two pairs all methods yield similar results, with the Y3 SOM riz having the best performance by a few percent for bins 0-2, and the SOMF $griz$ again having the best performance for bins 0-3.

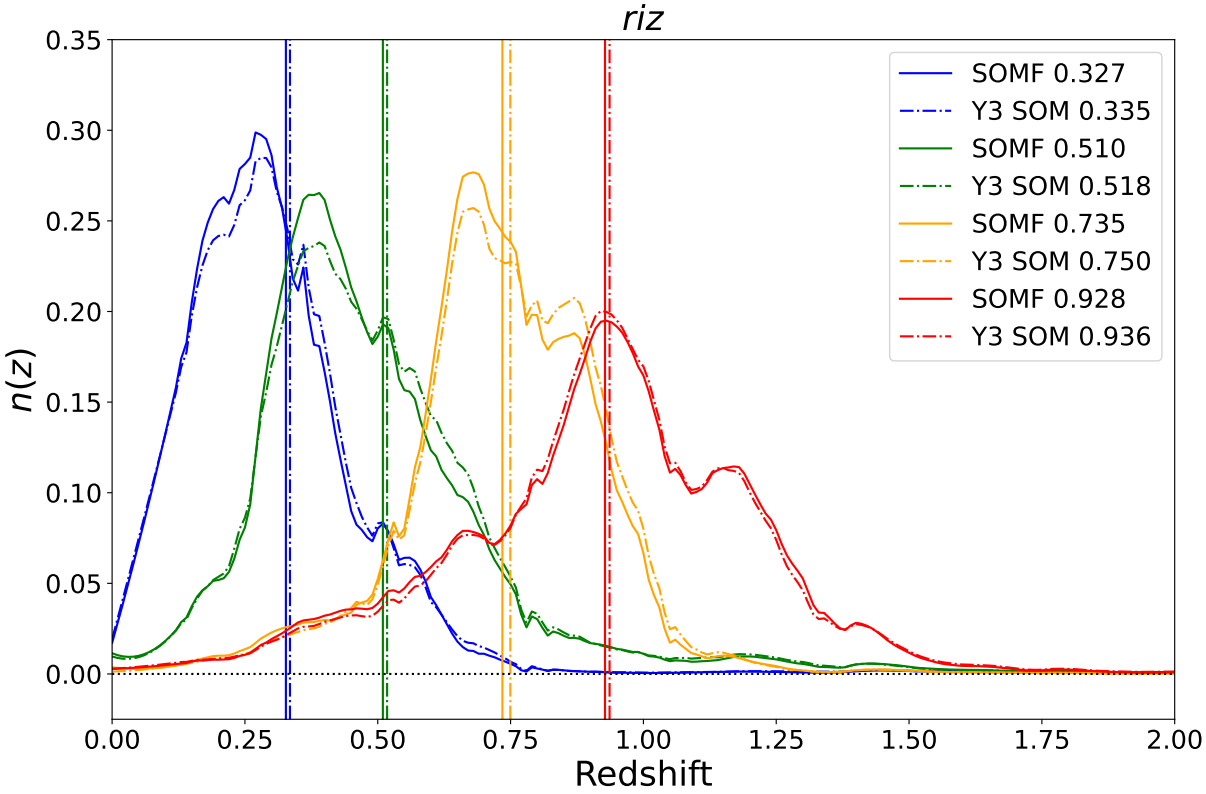

Figure 2.5: Photometric redshift distribution obtained from the *riz* bands, using the Y3 SOM (dot-dashed line) and the SOMF algorithm (filled line). The two methods show good agreement regarding the shape of each bin, and their mean redshifts. The SOMF method, however, presents better defined bins.

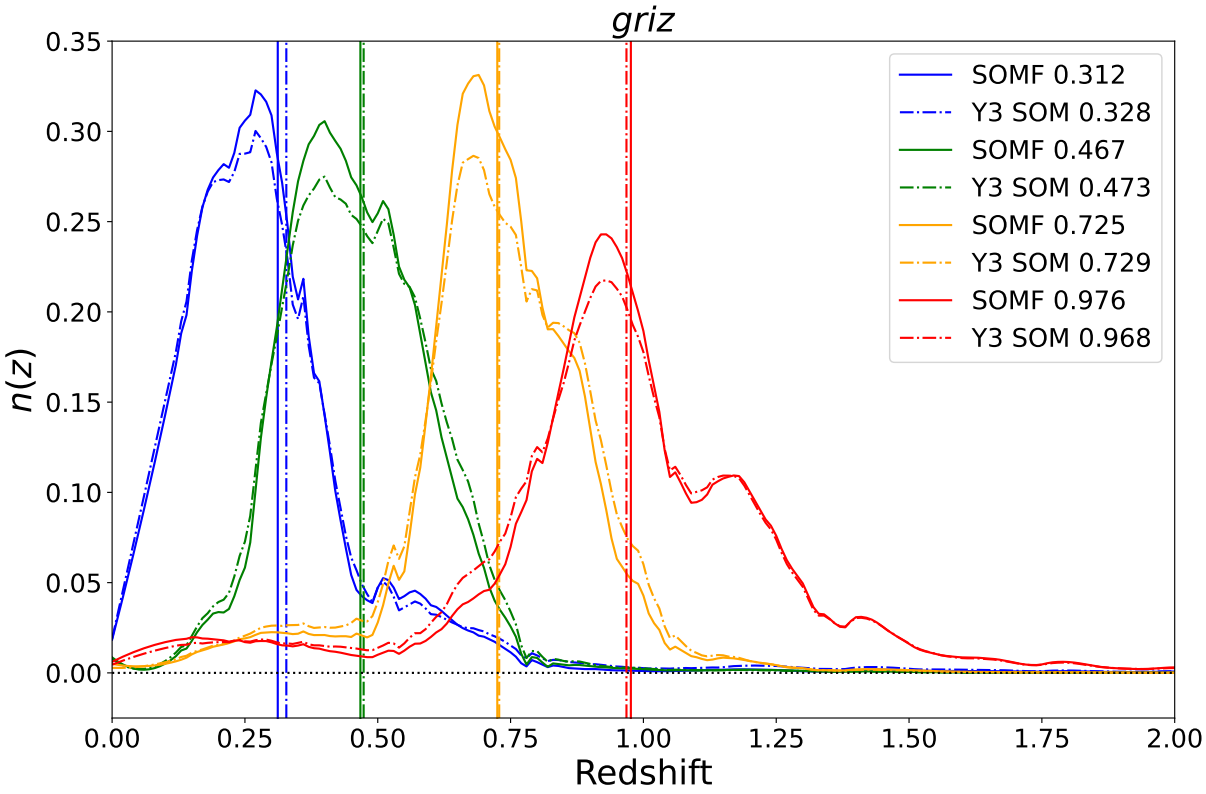

Figure 2.6: Photometric redshift distribution obtained from the *griz* bands, using the Y3 SOM (dot-dashed line) and the SOMF algorithm (filled line). The two methods show good agreement regarding the shape of each bin, and their mean redshifts, however the addition of the g-band further emphasizes the ability of SOMF to produce better defined bins.

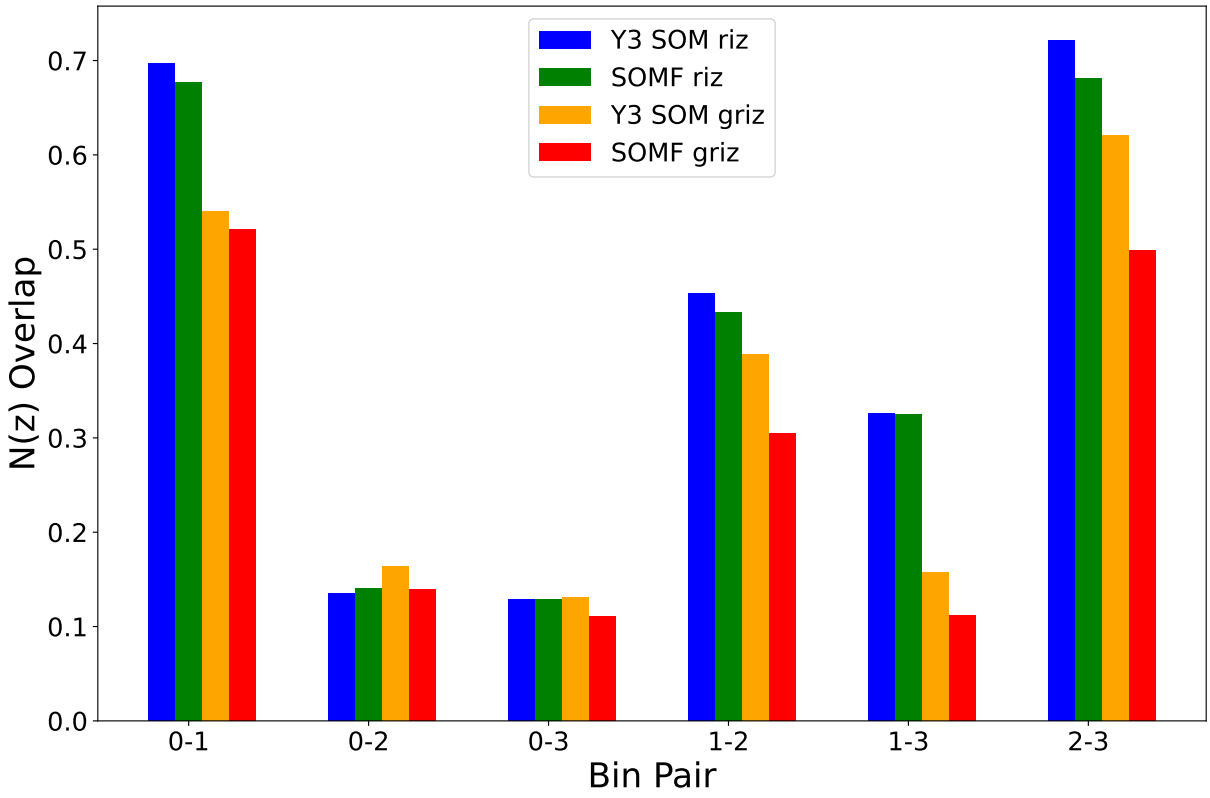

Figure 2.7: Redshfit bin overlap between bins $0-1$, $0-2$, $0-3$, $1-2$, $1-3$ and $2-3$ for each SOM recipe. The Y3 SOM *riz* is shown in blue, the SOMF *riz* in green, the Y3 SOM *griz* in yellow and the SOMF *griz* in red. We can see that all proposed modifications reduce the bin overlap with respect to the fiducial Y3 SOM using *riz* bands, with the best result obtained for SOMF *griz*.

	Method			
	Y3 SOM	SOMF	Y3 SOM	SOMF
Bin Pairs	*riz*	*riz*	*griz*	*griz*
0-1	0.6974	0.6767	0.5399	0.5207
0-2	0.1354	0.1407	0.1634	0.1393
0-3	0.1283	0.1283	0.1306	0.1112
1-2	0.4536	0.4329	0.3880	0.3053
1-3	0.3258	0.3253	0.1574	0.1113
2-3	0.7216	0.6812	0.6210	0.4986
	Overlap Reduction Relative to Y3 SOM *riz*			
0-1	-	3%	23%	25%
0-2	-	-4%	-21%	-3%
0-3	-	0%	-2%	13%
1-2	-	5%	14%	33%
1-3	-	0%	52%	66%
2-3	-	6%	14%	31%

Table 2.3: Amount of bin overlap between each redshift bin pair, for each method, together if the percentage overlap reduction with respect to the Y3 SOM *riz* (the fiducial method used in DES Y3).

In summary, using the same data and same methodology, described in Section 2.3.2, we can reduce the amount of bin overlap in our wide sample just by replacing the fiducial Y3 SOM algorithm used in the Y3 analysis (see Myles, Alarcon et al. 2021, Buchs et al. 2019), described in Section 2.3, by the SOMF algorithm (see Sánchez, Raveri, Alarcon & Bernstein 2020) described in Section 2.4, and adding the g-band information. By adding the g-band, the overlap between redshift bins undergoes significant improvement. The fiducial Y3 SOM already shows significant reduction in bin overlap when the g-band is added, but it is by combining the SOMF with the g-band information that we obtain a substantial reduction in bin overlap and the best defined redshift bins. In the next section, we show the impact of this reduction in bin overlap on the cosmological parameters.

2.6 Impact on cosmological parameters

In this Section we quantify how the changes in the SOMPZ method proposed in this paper impact the final cosmological constraints. In particular, we want to see how these changes impact the $S_8 - \Omega_{\rm m}$ plane, i.e., the main cosmology results for DES Y3.

Changing the redshift estimation of the source catalog impacts all the following steps in the cosmology estimation pipeline. In the SOMPZ method (see Section 2.3), the galaxy bin assignment is based on the wide SOM assignment. Therefore, when we train a new wide SOM, the galaxies are re-assigned and it is necessary to re-compute the two-point statistics measurements and the covariance matrix. In the case of the SOMF method with the *riz* bands, it is possible to perform all those steps using the Y3 data, and get a direct comparison between the Y3 SOM and the SOMF in the Y3 cosmology. In the cases when we add the g-band information, both for the Y3 SOM and the SOMF, it is not possible to carry the comparison all the way to the cosmological parameters, given that we don't have shape measurements for the g-band. Instead, we generated simulated data vectors, base on the Y3 cosmology, and compare the contours obtained in this simulated data.

Since in this paper we are exploring modifications on the method for redshift estimation for the weak lensing source catalog, we will be focusing on the cosmic shear measurement. Notice however, that the changes discussed here also impact galaxy-galaxy lensing and, naturally, the 3×2pt statistics.

2.6.1 Cosmological Constraints - Y3 Data

We tested the impact of replacing the SOM algorithm all the way from the SOM creation and assignment, to the cosmological parameter estimation. We use the redshift estimation schema and data described in Section 2.3.2, but replace the SOM algorithm outline in Section 2.3.1 (see also Buchs et al. 2019), with the one outlined in Section 2.4.1 (see also Sánchez, Raveri, Alarcon & Bernstein 2020).

Creating a new wide SOM and assigning the wide sample to it has a significant impact on bin assignments, influencing which galaxies are sorted into specific redshift bins. Consequently, to derive cosmological parameters, it becomes necessary to recalibrate various components of the analysis. This entails the reassessment of 2-point statistics, and the subsequent re-calculation of the covariance matrix. In essence, this process entails a complete reconstruction of the data-vector. Details on each step can be found in Appendix 2.C. Subsequently, we initiate a parameter estimation chain using the updated data-vector, adhering to the methodology outlined in the DES Y3 pipeline, as exhaustively expounded in Krause et al. (2021), and concisely summarized in Appendix 3.2. It's worth emphasizing that we follow the same methodology as Amon et al. 2022 and Secco, Samuroff et al. 2022.

Figure 2.8 compares the 1σ and 2σ contours in the $\Omega_m - \sigma_8$ plane. The blue contour uses the Y3 SOM algorithm, however, containing only the SOMPZ information when constraining the redshift (as opposed to the complete redshift information used in Y3 that contains SOMPZ + Clustering Redshifts (WZ) + Blending + Shear Ratios information, as detailed in Myles, Alarcon

et al. 2021). The green contour shows the chain for which the data vector was constructed with the redshift information from the SOMF algorithm.

The contours agree at the level of chain variance, and we do not observe any gain in constraining power on the cosmological parameters, due to the small reduction in bin overlap obtained when using the SOMF riz (see Figure 2.7). The marginalized mean S_8 and $\Omega_{\rm m}$ values in ΛCDM are:

$$\begin{aligned} S_8 &= 0.761^{+0.037}_{-0.027} \qquad \text{(Y3 SOM)} \\ \Omega_{\rm m} &= 0.298^{+0.046}_{-0.061} \qquad \text{(Y3 SOM)} \end{aligned}$$

and

$$\begin{aligned} S_8 &= 0.756^{+0.035}_{-0.030} \qquad \text{(SOMF)} \\ \Omega_{\rm m} &= 0.301^{+0.041}_{-0.066} \qquad \text{(SOMF)} \end{aligned}$$

where uncertainties are 68% confidence intervals. We can see that the values agree well within the uncertainty level, and the confidence intervals are also equivalent. Notice that the mean S_8 and $\Omega_{\rm m}$ for the Y3 SOM are not the same as the ones quoted in Amon et al. (2022) and Secco, Samuroff et al. (2022), given that here our chains includes redshifts estimated only with the SOMPZ method, but again they are in perfect agreement.

The good level of agreement of the two chains, and to they Y3 fiducial results, demonstrates the robustness of our method, one of the main results of this paper. This result, combined to the agreement in mean redshift and shape of the distribution, demonstrates that the SOMF algorithm is compatible with the SOMPZ pipeline, and robust against the cosmology results, validating it and making it a viable option for DES Year 6.

We emphasize that the improvements on cosmology due to the enhanced redshift methodology described in this paper could be more significant for a cosmic shear analysis more limited by redshift uncertainty than DES Y3.

2.6.2 Cosmological Constraints - Simulations

To assess the potential impact of incorporating g-band information into redshift estimation on cosmological parameters, we constructed simulated data-vectors based on the DES Y3 setup, as detailed in Appendix 3.3.1. Subsequently, we analyzed these simulations using the Y3 pipeline.

We generated simulated data for four distinct cases, each employing different methods: the Y3 SOM method with the riz bands, the SOMF method with the riz bands, the Y3 SOM method with the $griz$ bands, and the SOMF method with the $griz$ bands.

At the level of simulated data vectors we can already see differences. Switching from using the riz SOM $n(z)$s to $griz$ SOMF, we see a roughly 5-10% increase in the lensing ($\kappa\kappa$) signal in the uppermost $4, 4$ bin correlation. This is likely due to the small upwards shift in the mean, and the reduction in the weight of the low redshift tail. Much of the signal-to-noise of cosmic shear comes from these upper bin correlations, and so boosting the signal here is useful for optimising our cosmological constraint.

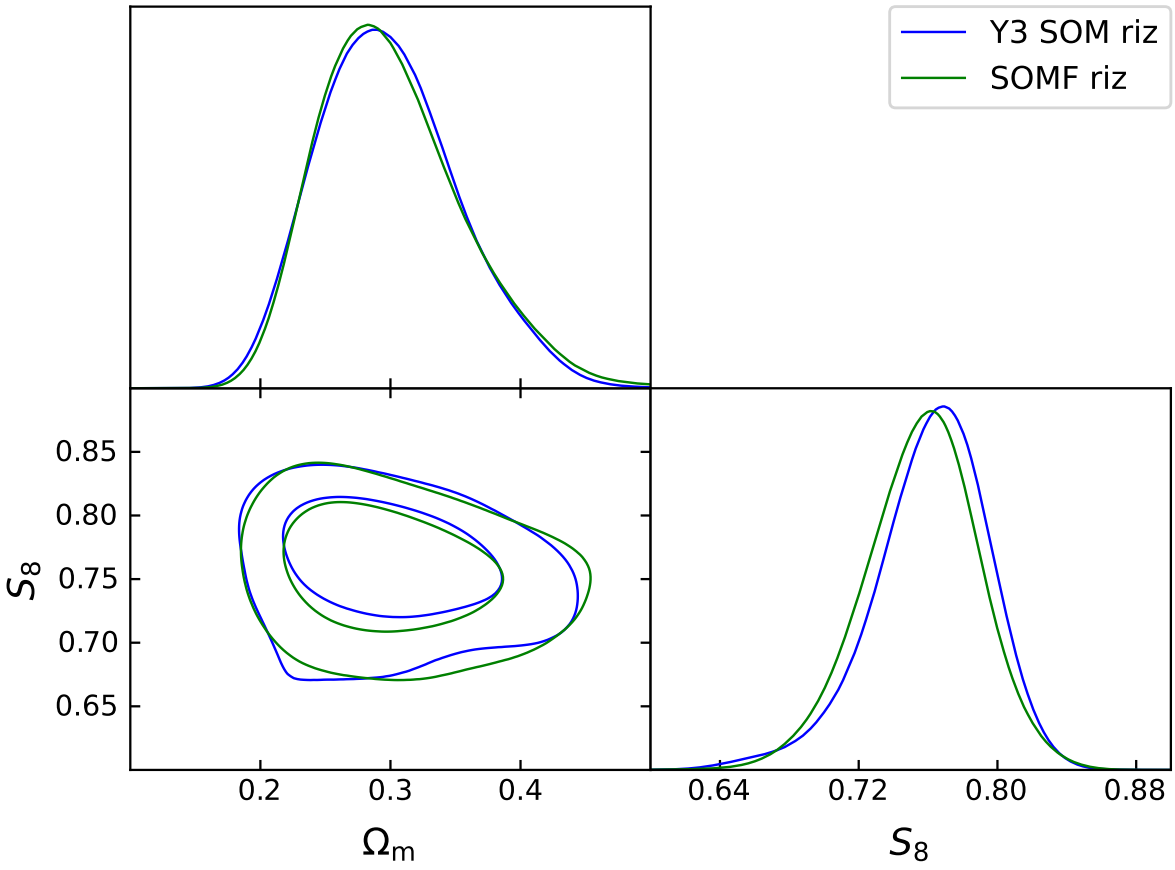

Figure 2.8: Cosmological constraints on the clustering amplitude, S_8 with the matter density, Ω_m in ΛCDM, using the DES Y3 data. The marginalised posterior contours (inner 68% and outer 95% confidence levels) are shown for the Y3 SOM in blue and SOMF in green.

We also see an overall reduction of the intrinsic alignment II contribution in the cross-bin correlations, as well as a decrease of the GI component in the auto bin pairs. This is again expected from a reduction in the width of the source bins (we can see this by considering, for example, Eq 16-17 of Secco, Samuroff et al. 2022). The cleaner separation of IA signals is helpful as it can break degeneracies and allow the data to constrain IAs more effectively.

In Figure 2.9, we present a comparison of the 1σ and 2σ contours in the $\Omega_m - \sigma_8$ plane for each case. Importantly, the cosmological parameters have been centered to zero in these plots, enabling us to concentrate on the gain in signal-to-noise (SN) or the reduction in errors relative to one another. This centered approach helps highlight how the various methods improve in comparison to their counterparts.

Upon examination, it is clear that the SOMF $griz$ stands out as the most effective method, showcasing superior constraining power to both Ω_m and σ_8. As such, it establishes itself as the optimal choice. Consequently, it is highly advisable to incorporate the SOMF $griz$ in the upcoming DES Year 6 analysis.

2.7 Conclusions

In this paper, we explore three modifications to the SOMPZ method (Buchs et al. (2019), Myles, Alarcon et al. 2021) employed in the DES Year 3 analysis: 1) changing the SOM algorithm; 2) including the photometry from the g-band; 3) adding the redshift information, when available, to train and assign galaxies to the SOM. Our goal is to optimize our redshift estimation pipeline for the DES Year 6 data, with especial focus on the weak lensing source galaxies. This is an important problem, given that those galaxies will be deeper and fainter compared to DES Y3, and we want to be able to treat them properly, minimizing cuts in our catalogs. That said, the findings of this paper are applicable to ensemble redshift estimation in general, and can be used as a guide for the lens redshifts in DES Year 6, and the redshift analysis of other surveys.

Using the DES Y3 weak lensing data, we tested each of the three modifications, and compared their impact relative to the Year 3 fiducial results. The main conclusions of our study are as follows:

- We showed that the SOMF successfully compresses the high-dimensional flux space into the redshift space in a smooth way, i.e., transitions between low- and high-redshift happen gradually. Given that Self-Organizing Map is a unsupervised clustering algorithm, neighbor cells should present similar properties. This property should definitely be observed in the flux space (the features the SOM was trained on, Figures 2.A.1 and 2.A.2), and the fact that its also present in redshift space, Figure 2.2, is evidence of the successful mapping of redshifts.

- By using the SOMF algorithm, tailored for the problem of photometric redshifts, we were able to reduce the standard deviation in the redshift distribution in each cell. That means that, within each cell of the SOM, we have a better estimation of the redshift of those galaxies.

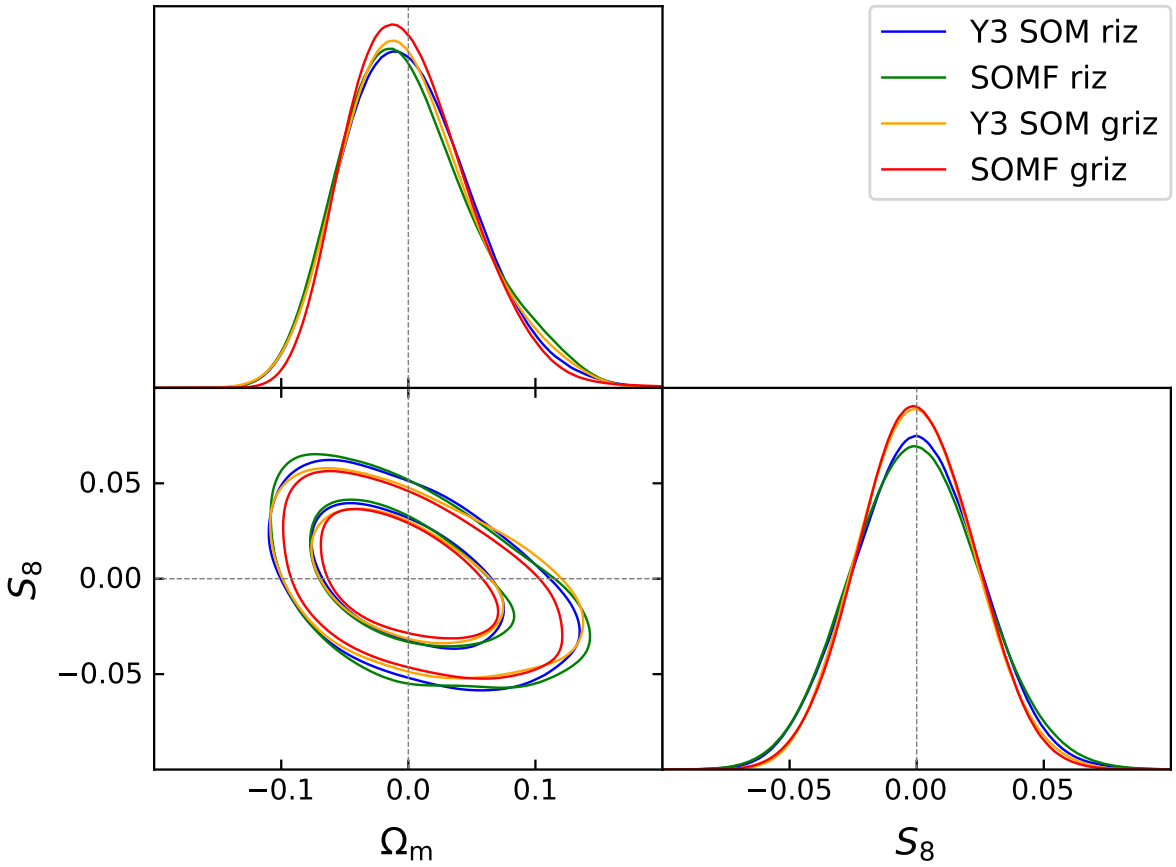

Figure 2.9: Cosmological constraints on the clustering amplitude, S_8 with the matter density, Ω_m in ΛCDM, using the simulated data described in Section 3.3.1, in order to include the g-band information. The marginalised posterior contours (inner 68% and outer 95% confidence levels) are shown for the Y3 SOM in blue and SOMF in green, for the *riz* bands, and Y3 SOM in yellow and SOMF in red for the *griz* bands.

- We are able to reduce bin-overlap even further, and therefore have better defined ensemble redshift distributions. Using the SOMF algorithm helps, however including an additional photometric band, the g-band, plays a major role in reducing the overlap between redshift bins.

- Two-point measurements and cosmology results are robust. Changing the galaxy assignment in the wide SOM affects all the following pipeline, meaning that all measurements need to be repeated. We tested the robustness of the 2pt cosmic shear measurements, and the final cosmology results, verifying that they are completely consistent with our Y3 findings.

- We considered adding the redshift measurements that we had available for galaxies in our spectroscopic sample as an additional feature to train the SOM and assign galaxies to it. This path did not lead to improvements in the method, as shown in Appendix 3.A, since the inclusion of the redshift feature seem to dominate over the other features, and create a very sparse SOM.

Photometric redshift estimation is an important topic in cosmology. There has been several proposals on how to improve redshift estimation, given the limitations imposed by photometric data. Self-Organizing Maps, in particular the SOMPZ pipeline proposed for DES Year 3 showed promising results, constraining the redshift of the weak lensing source galaxies to the 2% level. That was a ground-breaking work, that leverages on the deep fields high-quality photometric data to connect the spectroscopic information available for a small group of galaxies, to the main wide data set.

This paper sets out an recommendation for improvement of the redshift estimation pipeline, SOMPZ, for weak leasing source galaxies in DES Year 6. Although the fiducial SOMPZ method employed for Y3 is perfectly suitable for Y6 as well, by switching the Y3 SOM recipe to the SOMF and including the g-band photometry (even if only at the redshifts level) we can obtain even better redshift estimates for the Y3 wide sample. We expect these effects to be even more accentuated for Y6, given the increased depth of the wide sample, therefore the changes proposed here will have an even greater impact.

The implications of our results extend beyond the DES Y6 project. They provide a valuable foundation for the improvement and refinement of redshift characterization in future Stage IV surveys, such as those conducted by the Rubin Observatory. By reducing redshift bin overlap and enhancing the accuracy of redshift estimates, we are poised to unlock new possibilities for advancing our understanding of the universe's dark components, and to achieve more precise and robust cosmological parameter estimates in the years to come.

It is also worth emphasising that the code for the SOMPZ pipeline including the SOMF algorithm was further revised, simplified and documented, making it easier to use. Given its generality in the redshift estimation context, simplicity, and open source nature, we foresee the use of this method in future analyses, as a ensemble photometric redshift estimation pipeline option for many data sets.

Code Availability

The photometric redshift ensemble estimation code used in this work is publicly available at `https://github.com/AndresaCampos/sompz_y6`.

Acknowledgements

Andresa Campos thanks the support from the U.S. Department of Energy grant DE-SC0010118 and the NSF AI Institute: Physics of the Future, NSF PHY-2020295.

Appendix

2.A Magnitude and Colors - SOMF

Figures 2.A.1 and 2.A.2 provide visual representations of the *i*-band magnitude and colors for each cell within the wide and deep Self-Organizing Map (SOM), respectively. These maps were created using the SOMF algorithm, as outlined in Section 2.4.1. The SOM is designed to create a smooth map encompassing the entire parameter space derived from the training inputs. By drawing a comparison with Figure 2.2, we can appreciate how effectively the SOMF method maps the color-redshift relationship. This effectiveness is evident through the creation of a smooth redshift map that corresponds with the observed color patterns. The correlation between color and redshift in the SOM effectively illustrates its capacity to capture and represent this intricate relationship. Any abrupt differences observed between adjacent cells can be interpreted as indirect indications of potential degeneracies within the color-redshift relationship.

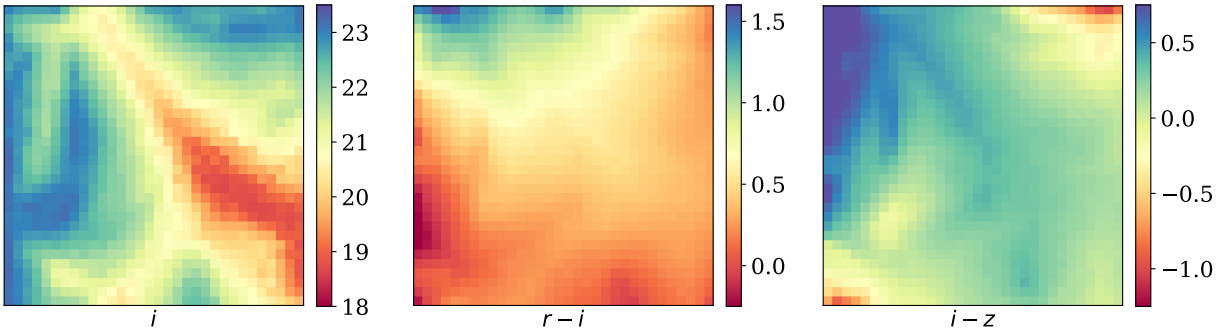

Figure 2.A.1: Wide Self-Organizing Map constructed using the SOMF algorithm and data from the *riz* bands. The visualization depicts the mean *i*-band magnitude (on the left), the mean *r* - *i* color (in the middle), and the mean *i* - *z* color (on the right) for each cell within the wide SOM.

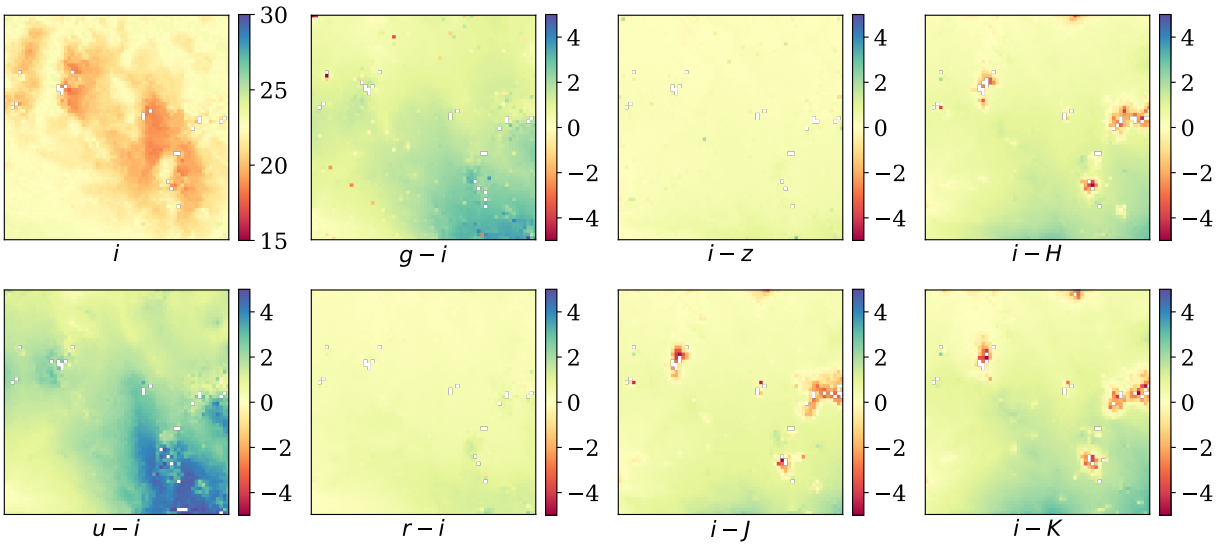

Figure 2.A.2: Deep field Self-Organizing Map constructed using the SOMF algorithm with data from the $ugrizJHK$ bands. In the upper left, we have the mean i-band magnitude for each cell within the deep SOM. Additionally, the various colors utilized in the deep SOM training are shown.

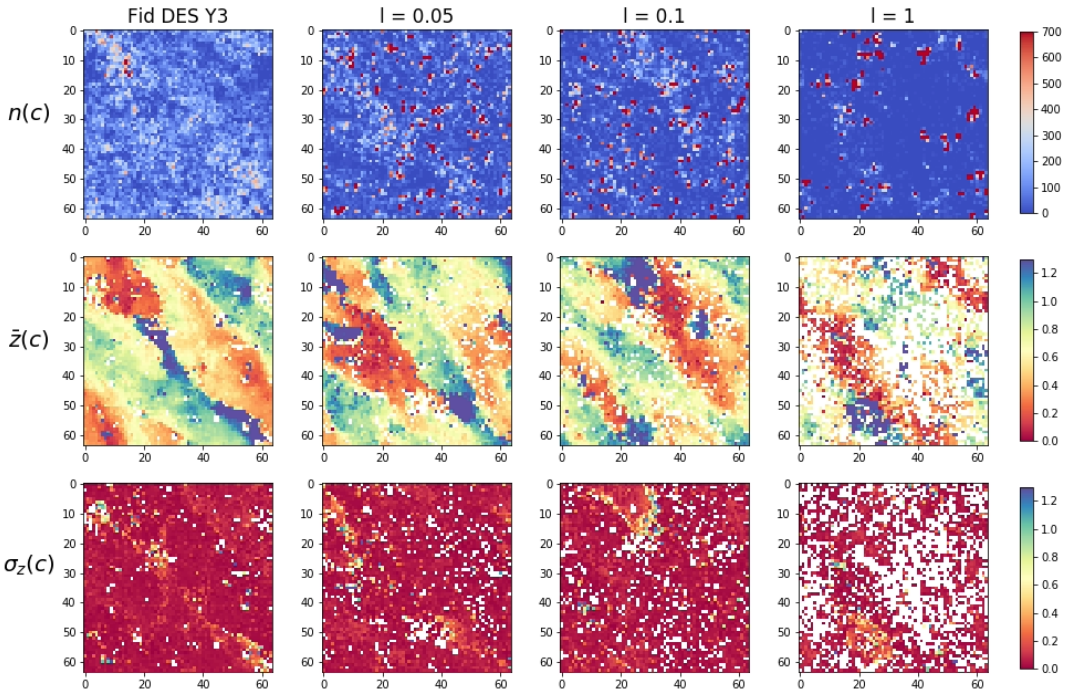

Figure 2.B.1: Adding redshift information in training and assigning deep SOM.

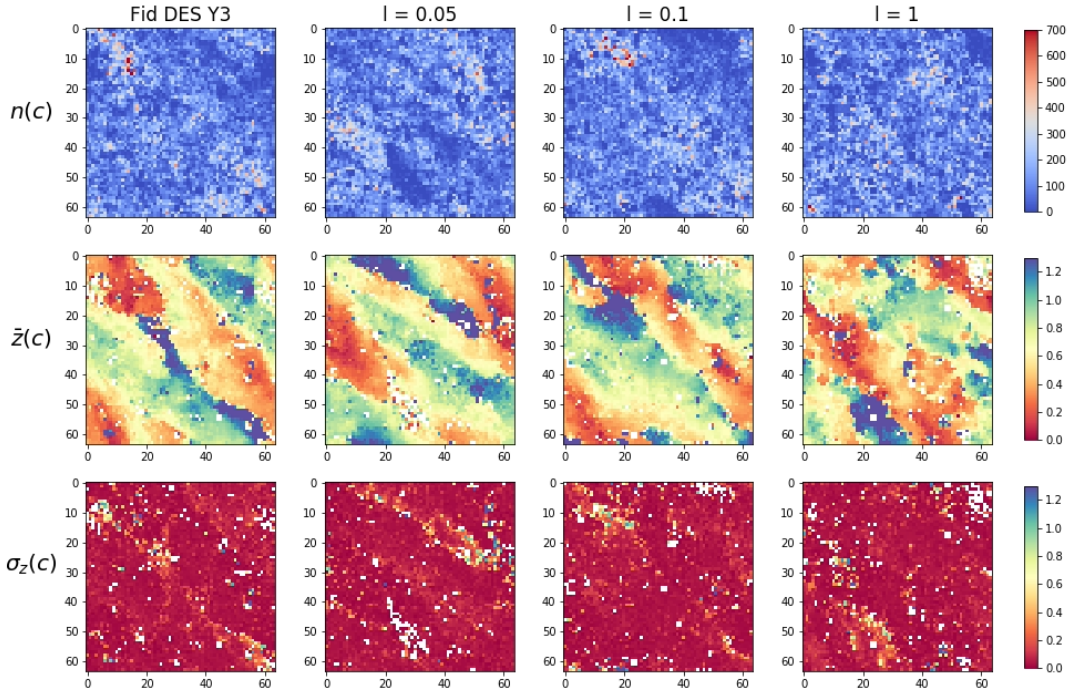

Figure 2.B.2: Adding redshift information in training deep SOM only.

2.B SOM-z

In this section, we delve into the details of incorporating redshift information, when available, as an extra feature in the training and assignment of the deep SOM used in the SOMPZ method. This is an unconventional way of applying an unsupervised learning method, such as the SOM, however we were motivated to investigate whether this approach would maximize the use of the available information, leading to enhancements in our ability to estimate redshifts. We applied this strategy to the Y3 SOM, described in Section 2.3. In light of what we observed, described in what follows, we did not try applying this strategy to the SOMF.

An interesting phenomenon arises when we include redshift information in the training and assignment process. The redshift sample galaxies tend to cluster in a few cells, leaving most cells sparsely occupied. This behavior can be observed in Figure 2.B (top), where we show the number of spectroscopic galaxies assigned to each cell in the deep SOM (top row), the mean redshift per cell (middle row), and the standard deviation in each cell (bottom row). The first column shows the DES Y3 SOM, without including the redshift information, and λ is a scaling factor for the contribution of redshift information relative to flux data, i.e., $\lambda = 0$ means no redshift contribution, while $\lambda = 1$ means the contribution of the redshift is the same as a flux. As the contribution of redshift information increases, the clustering effect intensifies.

This clustering of galaxies on the deep SOM negatively impacts the photometric redshift calibration. Figure 2.B.3 presents the wide data photometric distribution for the DES Y3 method and the variants with added redshift information. The solid line represents the DES Y3 $n(z)$, while the dot-dashed and dashed lines represent $\lambda = 0.05$ and $\lambda = 0.1$ respectively, and the dotted line represents the most "extreme" case, where $\lambda = 1$. As redshift information gains more weight, the $n(z)$ distribution in each redshift bin spreads further. This results in increased bin overlap, as illustrated in Figure 2.B.4, which is the opposite effect of what we are looking for.

Given that adding the redshift information the training and assigning phases of the deep SOM seems to impact negatively our $n(z)$ bins, we conducted a final test before abandoning the concept: adding the redshift information only during the training phase of the deep SOM. That still has a clustering effect upon the deep SOM, as we can see in Figure 2.B (bottom), but to a lesser degree. In this case, we observe that the $n(z)$ distribution in each bin, shown in Figure 2.B.5, and the bin overlap, shown in Figure 2.B.6, are very similar to those of the fiducial Y3 method, but still slightly worse given that the Y3 SOM still presents the smallest overlap.

Based on these findings, we infer that incorporating the redshift of individual galaxies as an additional feature alongside fluxes in the estimation of the $n(z)$ distribution using the SOMPZ method is not a viable approach.

2.C Cosmic Shear Measurement

The small distortions in the observed shapes of galaxies due to weak gravitational lensing by the intervening large scale structure of the Universe are called cosmic shear. Considering two redshift bins i and j, the shear correlation function estimator can be written in terms of a galaxy measured

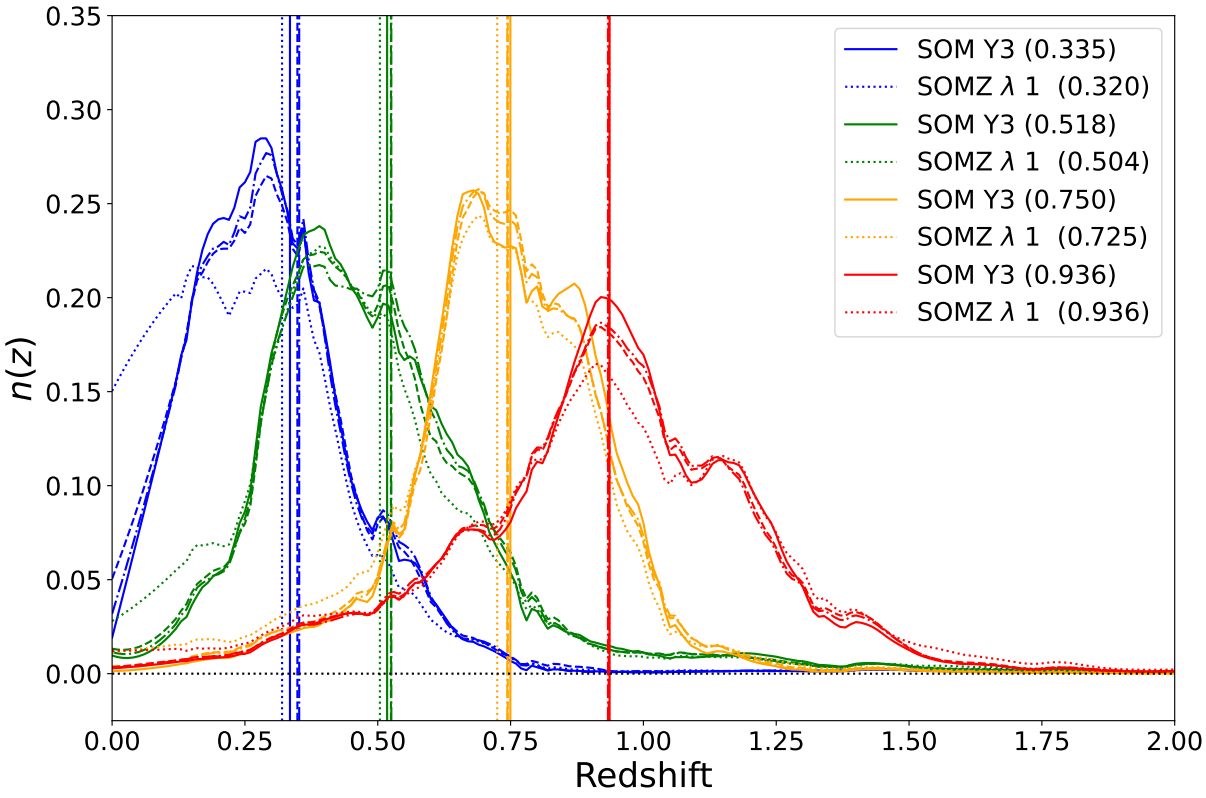

Figure 2.B.3: Comparison of the *n(z)* bins obtained using the fiducial DES Y3 SOM (solid line), and adding redshift in training and assigning deep SOM. The dotted line represents the most "extreme" case, where $\lambda = 1$ and the contribution of the redshift in training and assigning is the same as the fluxes, while the dashed and dot-dashed lines represent $\lambda = 0.1$ and $\lambda = 0.05$ respectively. The vertical lines are the mean redshift in each bin, shown in the legend for the fiducial method, or $\lambda = 0$, and the $\lambda = 1$ case.

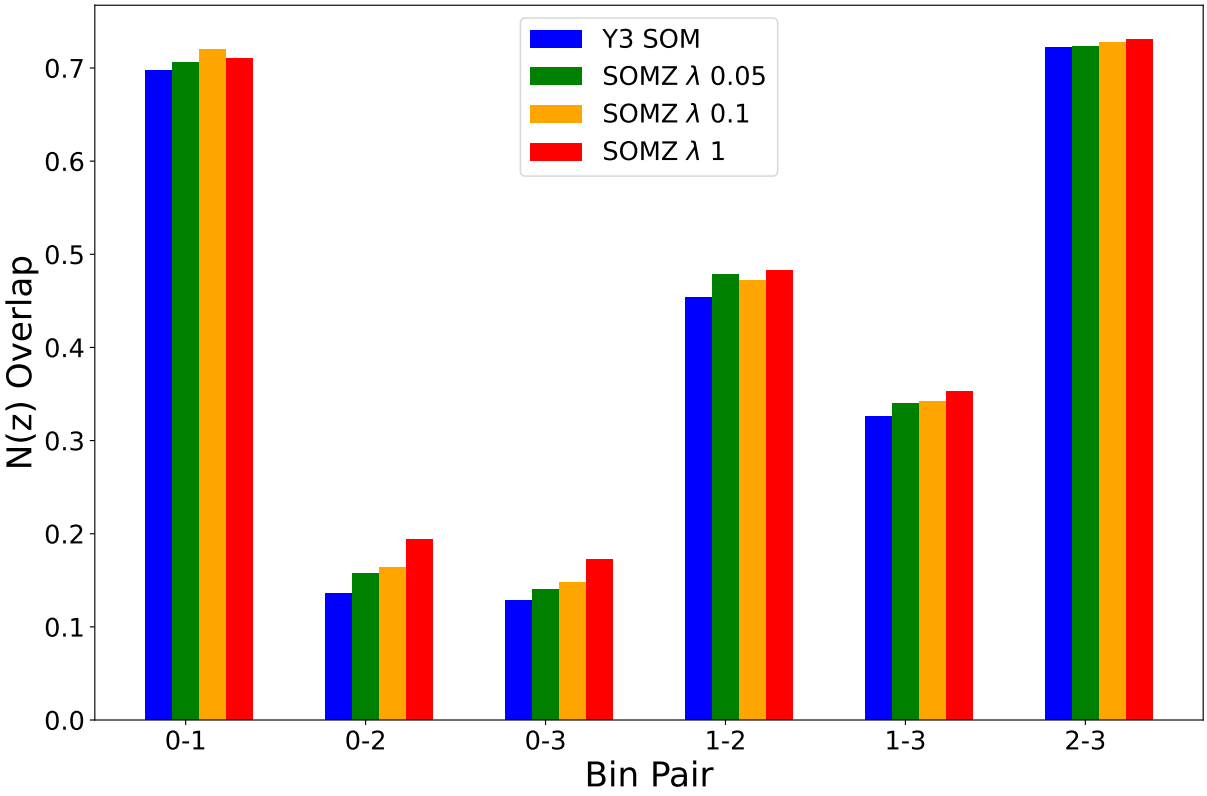

Figure 2.B.4: Redshfit bin overlap plot for fiducial DES Y3 (blue) and adding redshift in both training and assigning deep SOM. The bin overlap increases as the contribution of the redshift, represented by λ, increases. The green line represents $\lambda = 0.05$ or 5% contribution, the yellow $\lambda = 0.1$, contributing 10%, and the red line $\lambda = 1$, contributing the same as flux.

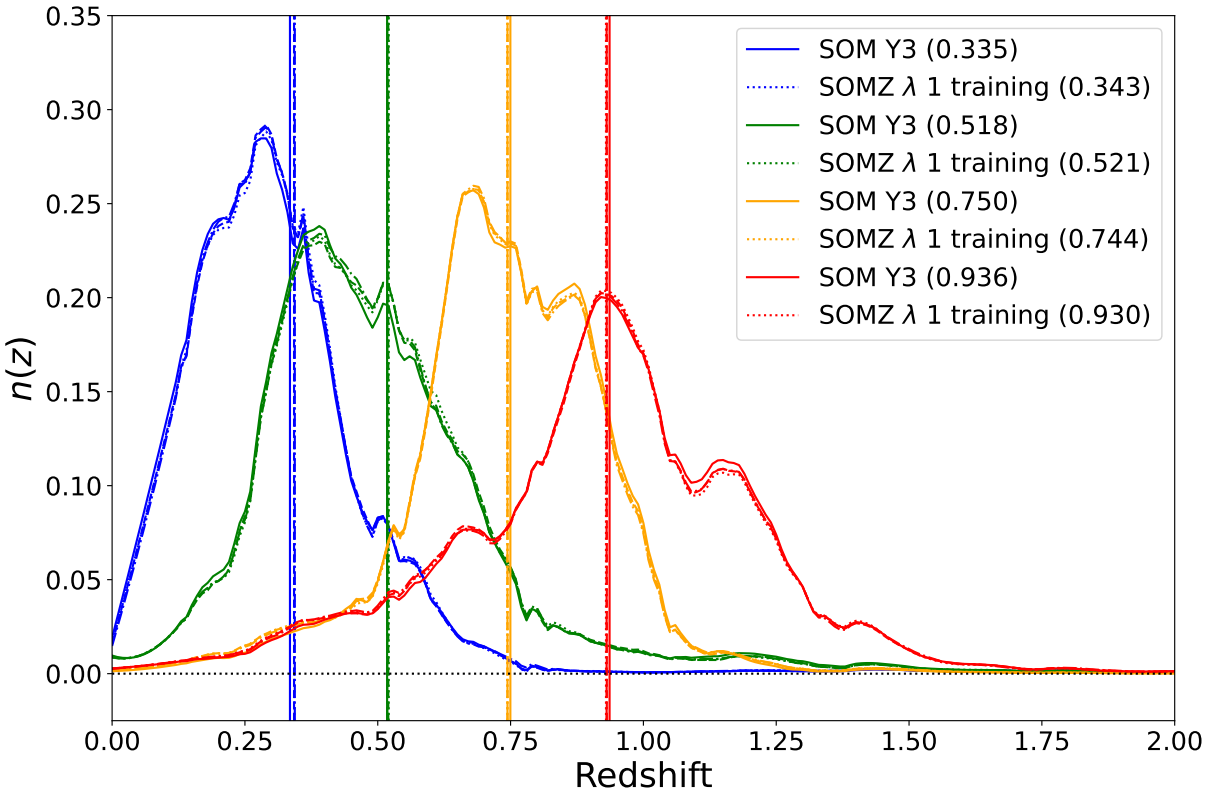

Figure 2.B.5: Comparison of the *n(z)* bins obtained using the fiducial DES Y3 SOM (solid line), and adding redshift only in the in training phase of the deep SOM. The dotted line represents the most "extreme" case, where $\lambda = 1$ and the contribution of the redshift in training and assigning is the same as the fluxes, while the dashed and dot-dashed lines represent $\lambda = 0.1$ and $\lambda = 0.05$ respectively. The vertical lines are the mean redshift in each bin, shown in the legend for the fiducial method, or $\lambda = 0$, and the $\lambda = 1$ case.

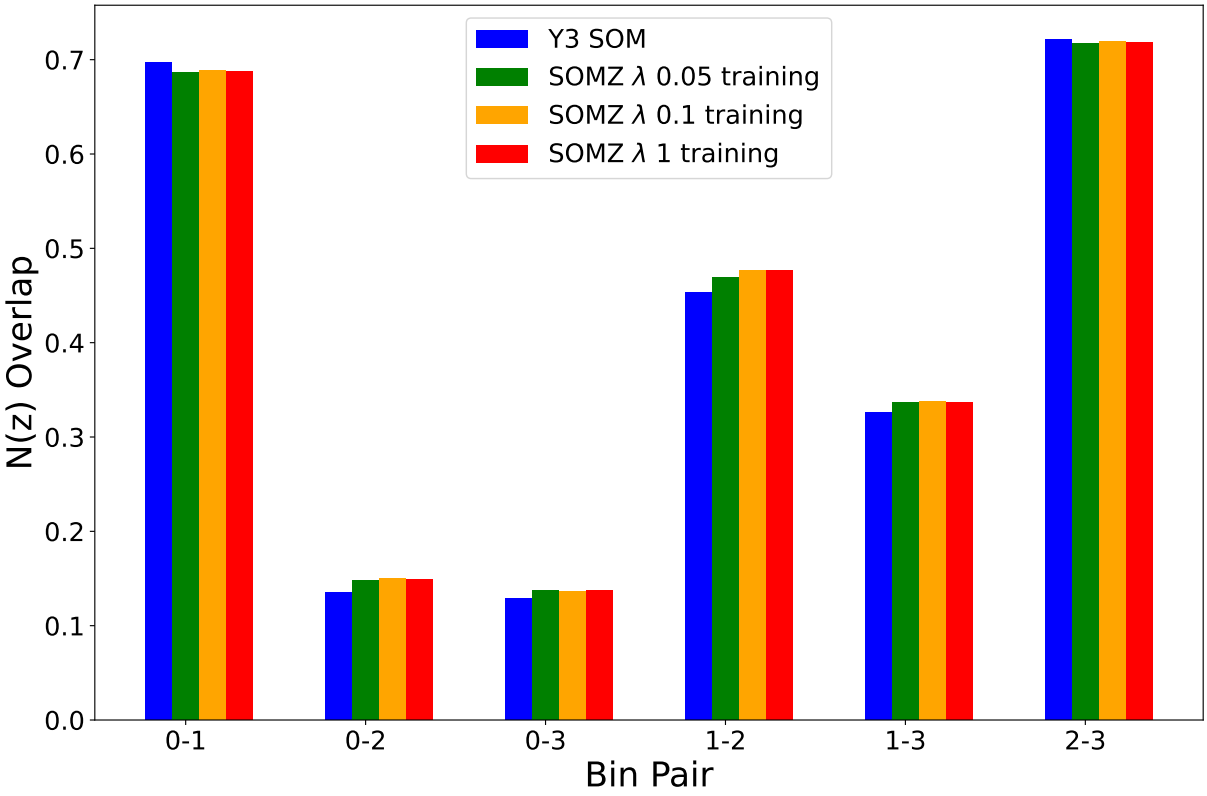

Figure 2.B.6: Redshfit bin overlap plot for fiducial DES Y3 (blue) and adding redshift in only training the deep SOM. The bin overlap increases as the contribution of the redshift, represented by λ, increases. The green line represents $\lambda = 0.05$ or 5% contribution, the yellow $\lambda = 0.1$, contributing 10%, and the red line $\lambda = 1$, contributing the same as flux.

tangential, ϵ_t, and radial, $\epsilon_\times$, ellipticities as

$$\xi_\pm^{ij}(\theta) = \langle \epsilon_t \epsilon_t \pm \epsilon_\times \epsilon_\times \rangle(\theta). \tag{2.21}$$

We can determine the shear-shear statistics by averaging over all galaxy pairs (a, b) separated by and angle θ

$$\xi_\pm^{ij}(\theta) = \frac{\sum_{ab} w_a w_b [\epsilon_t^i \epsilon_t^j \pm \epsilon_\times^i \epsilon_\times^j]}{\sum_{ab} w_a w_b R_a R_b} \tag{2.22}$$

where w represents the the per-galaxy inverse-variance weight, which is taken over galaxy pairs whose angular separation is within an interval $\Delta\theta$ around θ, and R is the shear response correction from Metacalibration.

The tomographic DES Y3 cosmic shear data vector, D, is computed using Equation 2.22 and the code TreeCorr[1]. It includes four auto-correlations and six cross-correlations between redshift bins for both positive and negative angular scales, spanning 2.5 to 250.0 arcmin. The impact of baryonic effects is mitigated by excluding small angular scales, leaving a total of 167 (60) data points for ξ_+ (ξ_-) correlations.

The covariance matrix, C, is a function of the redshift distributions, cosmological parameters and nuisance parameters. We assume a multivariate Gaussian distribution to model the statistical uncertainties in our cosmic shear data vector. The complete modelling of the disconnected 4-point function part of the covariance matrix is described in Friedrich et al. (2021). We compute the connected 4-point function part of the covariance matrix and the contribution from super-sample covariance using the public code CosmoCov[2] Fang et al. (2020), which is part of the CosmoLike framework Krause & Eifler (2017).

Following Amon et al. (2022), Secco, Samuroff et al. (2022) and previous cosmic shear analysis, we use a iteratively fixed covariance matrix. This means that we start with a set of fiducial input parameters, in our case we use the DES Y3 best fit parameters. Then the covariance is recomputed at the best fit from this first iteration, and the final chains are run.This update had negligible effects on the cosmic shear constraints that we present in this paper.

2.D Modelling and Analysis Choices

We carry out our analysis in the context of the flat ΛCDM cosmological model. The cosmological parameters are $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h_0, A_\mathrm{s}, n_\mathrm{s}, \Omega_\nu h^2\}$, where Ω_m is the density parameter for matter, and Ω_b the equivalent for baryons; h_0 is the dimensionless Hubble constant; A_s and n_s are the amplitude and slope of the primordial curvature power spectrum at a scale of $k = 0.05$ Mpc^{-1} respectively; and $\Omega_\nu h^2$ is the neutrino mass density parameter. We assume three degenerate massive neutrino species, following Krause et al. (2021).

[1] https://github.com/rmjarvis/TreeCorr

[2] https://github.com/CosmoLike/CosmoCov.

Modelling Cosmic Shear

For two redshift bins, i and j, the two-point cosmic shear correlations $\xi_{\pm}^{ij}(\theta)$ can be obtained by decomposing the convergence power spectrum $C_\kappa(\ell)$, at an angular wavenumber ℓ, into E- and B-mode components (Crittenden et al., 2002; Schneider et al., 2002)

$$\xi_{+}^{ij}(\theta) = \sum_{\ell} \frac{2\ell+1}{4\pi} G_{\ell}^{\pm}(\cos\theta) \left[C_{\kappa,EE}^{ij}(\ell) + C_{\kappa,BB}^{ij}(\ell) \right], \tag{2.23}$$

$$\xi_{-}^{ij}(\theta) = \sum_{\ell} \frac{2\ell+1}{4\pi} G_{\ell}^{\pm}(\cos\theta) \left[C_{\kappa,EE}^{ij}(\ell) - C_{\kappa,BB}^{ij}(\ell) \right], \tag{2.24}$$

where the functions $G_{\ell}^{\pm}(x)$ are calculated from Legendre polynomials $P_\ell(x)$ and averaged over angular bins (see Eqs. 19 and 20 in Krause et al. 2021).

The 2D convergence power spectrum $C_\kappa^{ij}(\ell)$ can be written in terms of the 3D matter power spectrum, assuming the Limber approximation (Limber, 1953; LoVerde & Afshordi, 2008), as:

$$C_\kappa^{ij}(\ell) = \int_0^{\chi(z_{\rm max})} \mathrm{d}\chi \frac{W^i(\chi) W^j(\chi)}{\chi^2} P_\delta\left(\frac{\ell+0.5}{\chi}, z(\chi) \right), \tag{2.25}$$

where $P_\delta(k,z)$ is the nonlinear matter power spectrum and the lensing weight is:

$$W^i(\chi) = \frac{3H_0^2\Omega_{\rm m}}{2c^2} \frac{\chi}{a(\chi)} \int_\chi^{\chi_{\rm H}} \mathrm{d}\chi'\, n^i\left(z(\chi')\right) \frac{\mathrm{d}z}{\mathrm{d}\chi'} \frac{\chi'-\chi}{\chi'}, \tag{2.26}$$

with the source galaxy redshift distribution $n^i(z)$ normalised to integrate to 1, and $\chi_{\rm H}$ the horizon distance. We follow Krause et al. (2021), and model P_δ using a combination of CAMB (Lewis et al., 2000) for the linear part, and HaloFit (Takahashi et al., 2012) for nonlinear modifications. As highlighted in Amon et al. (2022) and Secco, Samuroff et al. (2022), the impact of higher order contributions to the observed two-point statistics is verified to be negligible for the scales covered in this work.

Nuisance Parameters & Scale Cuts

Our setup matches the fiducial choices of the DES Y3 cosmic shear analysis. The only significant difference is that, for the sake of simplicity, we choose not to use the additional shear ratio likelihood included by Amon et al. (2022); Secco, Samuroff et al. (2022) (a similar decision was made for validating the analysis choices pre-unblinding; see Krause et al. 2021). As a result, our model space is slightly smaller, since we do not need to vary parameters for galaxy bias or lens photo$-z$ error. The corresponding parameters and their priors are shown in Table 3.A.1. Note that these are identical to the priors used in the Y3 analysis. We also adopt the fiducial DES Y3 cosmic shear scale cuts (see Krause et al. 2021 for an explanation of how these were derived).

Table 2.D.1: A summary of the central values and priors used in our analysis. The top seven rows are cosmological parameters, while those in the lower sections are nuisance parameters corresponding to astrophysics and data calibration. Priors are either uniform (U) or normally-distributed, $\mathcal{N}(\mu, \sigma)$.

Parameter	**Fiducial Value**	**Prior**
Cosmological Parameters		
Ω_m	0.29	$\mathrm{U}[0.1, 0.9]$
Ω_b	0.052	$\mathrm{U}[0.03, 0.07]$
h	0.75	$\mathrm{U}[0.55, 0.91]$
A_s	2.38×10^{-9}	$\mathrm{U}[0.5, 5.0] \times 10^{-9}$
n_s	0.99	$\mathrm{U}[0.87, 1.07]$
$\Omega_\nu h^2$	0.00053	$\mathrm{U}[0.6, 6.44] \times 10^{-3}$
Calibration Parameters		
m_1	0.0	$\mathcal{N}(0.0, 0.0059)$
m_2	0.0	$\mathcal{N}(0.0, 0.0042)$
m_3	0.0	$\mathcal{N}(0.0, 0.0054)$
m_4	0.0	$\mathcal{N}(0.0, 0.0072)$
Δz_1	0.0	$\mathcal{N}(0.0, 0.018)$
Δz_2	0.0	$\mathcal{N}(0.0, 0.015)$
Δz_3	0.0	$\mathcal{N}(0.0, 0.011)$
Δz_4	0.0	$\mathcal{N}(0.0, 0.017)$
Intrinsic Alignment Parameters		
A_1	0.7	$\mathrm{U}[-5, 5]$
A_2	-1.36	$\mathrm{U}[-5, 5]$
η_1	-1.7	$\mathrm{U}[-5, 5]$
η_2	-2.5	$\mathrm{U}[-5, 5]$
b_TA	1.0	$\mathrm{U}[0, 2]$

Generating Mock Data

In this section, we outline the process of generating mock data, which serves as a means to assess the impact of including the g-band in the DES Y3 setup. For a set of input parameters, we generate four noiseless DES Y3-like cosmic shear data vectors denoted as $\boldsymbol{D}$. These data vectors are produced using the theoretical pipeline described in Section 3.2 and are centered around the central values outlined in Table 3.A.1.

All four data vectors share the same input flat ΛCDM cosmological model, with parameters set as follows: $\Omega_{\rm m} = 0.29$, $A_{\rm s} = 2.38 \times 10^{-9}$, $\Omega_{\rm b} = 0.052$, $h = 0.75$, $n_{\rm s} = 0.99$, and $\Omega_\nu h^2 = 0.00053$. This configuration corresponds to $\sigma_8 = 0.79$ and $S_8 = 0.77$, where $S_8 \equiv \sigma_8\sqrt{\Omega_{\rm m}/0.3}$. However, each data vector is distinct in terms of the redshift distribution of the source galaxies, determined using one of the methods employed in this study: 1) One data vector utilizes the redshift distribution obtained by the Y3 SOM, using the *riz* bands. 2) Another data vector adopts the redshift distribution obtained by the SOMF method, utilizing the *riz* bands. 3) A third data vector relies on the redshift distribution derived from the Y3 SOM, employing the *griz* bands. 4) The final data vector is constructed with the redshift distribution acquired through the SOMF method, employing the *griz* bands. Our analysis framework and mock data generation follow the choices made in the DES Y3, ensuring that our assessments are consistent with the established DES Year 3 standards.

Bayesian Inference

For the purpose of parameter estimation, the *likelihood function* of the data vector, D, given the model, T, characterized by parameters, $\mathbf{p}$, can be represented as $\mathcal{L}(D|\mathbf{p})$. This probability distribution is presumed to follow a multivariate Gaussian distribution

$$\ln \mathcal{L}(D|\mathbf{p}) = -\frac{1}{2}\sum_{ij}\Big(D_i - T_i(\mathbf{p})\Big)[C]^{-1}_{ij}\Big(D_j - T_j(\mathbf{p})\Big) \tag{2.27}$$

D_i represents the ith component within the data vector $\xi_\pm$, together with its covariance matrix, C (see Section 2.C). Initially, this vector incorporates 20 angular data points, each spanning across the intersections of 4 redshift bins and 2 correlation functions, leading to a total of 227 data points after constraining the angular scales. The corresponding theoretical predictions for these statistical quantities, represented as $T_i(\mathbf{p})$, are elaborated upon in this section. The Bayesian *posterior probability* distributions of the cosmological parameters, denoted as $\mathcal{P}(\mathbf{p}|D)$, are derived by combining the likelihood with the *prior probabilities*, $P(\mathbf{p})$, as outlined in Table 3.A.1, following the principles of Bayes' theorem

$$\mathcal{P}(\mathbf{p}|D) = \frac{P(\mathbf{p})\mathcal{L}(D|\mathbf{p})}{P(D)} \tag{2.28}$$

where $P(D)$ is the *evidence* of the data.

The posterior distribution is sampled using the Polychord (Handley et al., 2015a,b). The analysis framework is based on CosmoSIS (Zuntz et al., 2015), a modular tool for estimating

cosmological parameters. We use the fiducial sampler settings (500 live points, tolerance 0.01) that have been verified to showcase the precision of the posterior distributions and Bayesian evidence estimations (as discussed in Lemos et al. 2022).

Chapter 3

An empirical approach to model selection: weak lensing and intrinsic alignments

Abstract

In cosmology, we routinely choose between models to describe our data, and can incur biases due to insufficient models or lose constraining power with overly complex models. In this paper we propose an empirical approach to model selection that explicitly balances parameter bias against model complexity. Our method uses synthetic data to calibrate the relation between bias and the χ^2 difference between models. This allows us to interpret χ^2 values obtained from real data (even if catalogues are blinded) and choose a model accordingly. We apply our method to the problem of intrinsic alignments – one of the most significant weak lensing systematics, and a major contributor to the error budget in modern lensing surveys. Specifically, we consider the example of the Dark Energy Survey Year 3 (DES Y3), and compare the commonly used nonlinear alignment (NLA) and tidal alignment & tidal torque (TATT) models. The models are calibrated against bias in the $\Omega_{\rm m} - S_8$ plane. Once noise is accounted for, we find that it is possible to set a threshold $\Delta\chi^2$ that guarantees an analysis using NLA is unbiased at some specified level $N\sigma$ and confidence level. By contrast, we find that theoretically defined thresholds (based on, e.g., $p-$values for χ^2) tend to be overly optimistic, and do not reliably rule out cosmological biases up to $\sim 1 - 2\sigma$. Considering the real DES Y3 cosmic shear results, based on the reported difference in χ^2 from NLA and TATT analyses, we find a roughly 30% chance that were NLA to be the fiducial model, the results would be biased (in the $\Omega_{\rm m} - S_8$ plane) by more than 0.3σ. More broadly, the method we propose here is simple and general, and requires a relatively low level of resources. We foresee applications to future analyses as a model selection tool in many contexts.

methods: statistical – cosmology: observations – cosmological parameters – gravitational lensing: weak

3.1 Introduction

Modern cosmology is an increasingly high-dimensional problem. Although the standard cosmological model itself is relatively simple, containing only five or six free parameters, it cannot, in general, be constrained in isolation. One must rely on measurements on real data, which can contain any number of additional features resulting from non-cosmological processes. It is necessary to include models for such systematics in any cosmological inference, and to marginalise over their parameters. Contemporary weak lensing analyses (see, e.g., Heymans et al. 2013; DES Collaboration 2016a; Jee et al. 2016; Hildebrandt et al. 2017; Troxel et al. 2018; Hikage et al. 2019; Hamana et al. 2020; Asgari et al. 2021; Loureiro et al. 2022; Amon et al. 2022; Secco, Samuroff et al. 2022; Doux et al. 2022) typically have around $15 - 30$ free parameters, the majority of which are related to measurement uncertainties. This picture is unlikely to change in the coming years. Indeed, as we move into the era of Stage IV surveys (Ivezić et al. 2019; Spergel et al. 2015; Laureijs et al. 2011), the unprecedented statistical power of these new data sets carries an increasing sensitivity to systematics.

Some systematic uncertainties can be modelled pretty accurately given our prior knowledge of their nature; for instance PSF modelling error (Jarvis et al., 2021) and shear measurement biases (Heymans et al., 2006; Bridle et al., 2010; Mandelbaum et al., 2015). In most cases, however, there is a relative lack of prior knowledge about the magnitude and/or scale dependence of the effects being modelled. Some examples include the impact of baryonic feedback (Osato et al., 2015; Chen et al., 2023; Tröster et al., 2022), nonlinear structure formation (and the impact of neutrinos on it; Saito et al. 2008; Bird et al. 2012; Mead et al. 2021; Knabenhans et al. 2021) and galaxy bias (Desjacques et al., 2018; Simon & Hilbert, 2018; Pandey et al., 2020). Here, there is clearly an argument for using the most sophisticated (physically motivated) model available. This is the safest way to avoid bias due to model insufficiency. That said, extra free parameters do potentially carry a cost in terms of constraining power. They can also worsen projection effects, which complicate the interpretation of projected parameter constraints (see Joachimi et al. 2021a; Krause et al. 2021). The ideal approach, then, would be to select a model that balances the two: complex enough to avoid bias, but not more complex than is needed to describe the data.

Model selection methods are widely used in cosmology, often seeking to answer the question of whether introducing new parameters to cosmological models is justified by the data. Some of the most common tools for this are χ^2 tests, the Akaike Information Criterion (AIC), the Bayesian Information Criterion (BIC), the Deviance Information Criterion (DIC) and Bayes ratios (see e.g. Liddle et al. 2006a; Liddle 2007; Trotta 2007; Trotta 2008; Kerscher & Weller 2019). A characteristic of how all these statistics have been used is that they are interpreted using threshold statistical values, derived in terms of the theoretical properties of the model, e.g. the Jeffreys scale for Bayes ratios. They have also been most commonly applied to compare how well cosmological models fit the data post-analysis, rather than actively being used to select elements of the analysis in the blinded stages. By contrast, the process for choosing the fiducial model for an analysis typically does not make use of model comparison statistics at all. Rather, we tend to rely on generating and analysing simulations (either analytic or numerical) containing various forms of unmodelled systematics (e.g., Krause et al. 2017, 2021; Joachimi et al. 2021a). This approach

works, but does heavily depend on the ability to create realistic mocks. It is also important to notice that any model selection method will typically have a number of subjective choices built into them, e.g., whether to compare data vectors or perform likelihood inference, and what cutoff to use for decision-making. This is also true, to an extent, for the method we will present in this paper. That said, our method has the feature that the decision-making happens in well-defined places and has a well-defined interpretation connected to parameter biases (e.g. selecting a tolerable bias level and a confidence interval), as we will see in the following sections.

One of the most significant sources of systematic uncertainty in weak lensing is an effect known as intrinsic alignment (IA; Troxel & Ishak 2015; Joachimi et al. 2015; Kirk et al. 2015; Kiessling et al. 2015). IAs are coherent galaxy shape alignments that are not purely due to lensing, but rather to the interactions with the local and large-scale gravitational field. Although in essence an astrophysical effect, IA correlations appear on much the same angular scales as cosmological ones, and it can be very difficult to disentangle the two. They are not universal, in the sense that they depend significantly on the particular galaxy sample (colour, luminosity, satellite fraction and redshift distribution; e.g. Johnston et al. 2019), and also the details of the shape measurement (Singh & Mandelbaum, 2016). To add to the problem, unlike, for example, photometric redshift error or shear bias, one cannot easily derive tight priors on IAs using simulations or external data. Some physically-motivated IA models that have been developed in the last two decades include the linear alignment (LA) model (Catelan et al., 2001; Hirata & Seljak, 2004; Hirata & Seljak, 2010) which, as the name suggests, assumes a linear relationship between galaxy shapes and the local tidal field; an empirical modification of this, known as the nonlinear alignment model (NLA; Hirata et al. 2007; Bridle & King 2007), which is now one of the most common IA models in contemporary weak lensing; and in recent years, the tidal alignment and tidal torquing model (TATT; Blazek et al. 2015, 2019), which has provided a slightly more complex alternative to NLA. Based on perturbation theory, TATT includes additional terms that are quadratic in the tidal field, intended to encapsulate the processes driving IAs in spiral galaxies, and also additional terms that are designed to enable better IA modelling on smaller (but still 2-halo) scales.

In this paper we propose a new model selection method, which uses the real data. The general idea is to run two competing models on the blinded data, and compare them using statistical metrics. Here we explore two convenient metrics: the difference in the best χ^2 per degree of freedom, $\Delta\chi^2_{(\mathrm{df})} = \Delta\chi^2/\Delta\mathrm{df}$, and the Bayes ratio R. We show that, for the method we are proposing, $\Delta\chi^2_{(\mathrm{df})}$ is a very useful metric to perform model selection (R is less so, for reasons discussed in Section 3.5.2 and Appendix 3.C). To allow us to interpret the $\Delta\chi^2_{(\mathrm{df})}$, we use simulated data to calibrate its relation to biases in parameter space due to model insufficiency. It is this process, of running a set of simulations and measuring parameter bias as a function of observable metrics that we refer to as "calibrating the bias-metric relation" in the following sections. The full details on how to perform this calibration are outlined in Section 3.4. This approach can, in principle, be applied to any type of data and/or systematics. Some examples of interesting use cases in cosmology are choosing models for galaxy bias, the nonlinear $P(k)$ and baryonic feedback, as well as extensions to ΛCDM and modified gravity. However, in what follows, we apply the method to the specific scenario of choosing an intrinsic alignment model for a cosmic shear analysis.

The paper is structured as follows: we describe the theoretical modelling of the cosmic shear two-point data vector in Section 3.2. In Section 3.3 we describe how the synthetic cosmic shear data is generated, including the choice of IA scenarios. The ingredients and steps for the model selection method are described in Section 3.4. Our results when applying our method to the problem of IAs in the Dark Energy Survey Year 3 (DES Y3) are presented in Section 4.5. Finally, in Section 4.7 we summarise our findings and their significance in the context of the field.

3.2 Theory & Modelling

We carry out our analysis in the context of the flat ΛCDM cosmological model. The cosmological parameters are $\{\Omega_\mathrm{m}, \Omega_\mathrm{b}, h_0, A_\mathrm{s}, n_\mathrm{s}, \Omega_\nu h^2\}$, where Ω_m is the density parameter for matter, and Ω_b the equivalent for baryons; h_0 is the dimensionless Hubble constant; A_s and n_s are the amplitude and slope of the primordial curvature power spectrum at a scale of $k = 0.05$ Mpc^{-1} respectively; and $\Omega_\nu h^2$ is the neutrino mass density parameter. We assume three degenerate massive neutrino species, following Krause et al. (2021). We discuss the nuisance parameters of our analysis in the following sections. Prior choices are further described in Appendix 3.A.

3.2.1 Modelling Cosmic Shear

The impact of gravitational lensing along a particular line of sight is determined by two quantities, known as convergence and shear. The *convergence* κ term of the weak lensing transformation describes how much a galaxy on a particular line of sight is distorted due to intervening large scale structure. It is defined as the weighted mass overdensity δ, integrated along the line of sight to the distance of the source χ_s:

$$\kappa(\boldsymbol{\theta}) = \int_0^{\chi_\mathrm{s}} \mathrm{d}\chi\, W(\chi)\delta(\boldsymbol{\theta}, \chi), \tag{3.1}$$

where $\boldsymbol{\theta}$ is the angular position at which the source is observed. The kernel $W(\chi)$, defined in Eq. (3.5), is sensitive to the relative distances of the source and the lens. It is this geometrical term that makes cosmic shear sensitive to the expansion history of the Universe.

The two-point cosmic shear correlations $\xi_\pm(\theta)$ are obtained by decomposing κ into E- and B-mode components (Crittenden et al., 2002; Schneider et al., 2002). For two redshift bins i and j, they can be written in terms of the convergence power spectrum $C_\kappa(\ell)$ at an angular wavenumber ℓ as

$$\xi_+^{ij}(\theta) = \sum_\ell \frac{2\ell+1}{4\pi} G_\ell^\pm(\cos\theta) \left[C_{\kappa,EE}^{ij}(\ell) + C_{\kappa,BB}^{ij}(\ell)\right], \tag{3.2}$$

$$\xi_-^{ij}(\theta) = \sum_\ell \frac{2\ell+1}{4\pi} G_\ell^\pm(\cos\theta) \left[C_{\kappa,EE}^{ij}(\ell) - C_{\kappa,BB}^{ij}(\ell)\right], \tag{3.3}$$

where the functions $G_\ell^\pm(x)$ are calculated from Legendre polynomials $P_\ell(x)$ and averaged over angular bins (see Eqs. 19 and 20 in Krause et al. 2021).

Assuming the Limber approximation (Limber, 1953; LoVerde & Afshordi, 2008), the 2D convergence power spectrum $C_\kappa^{ij}(\ell)$ is related to the 3D matter power spectrum as:

$$C_\kappa^{ij}(\ell) = \int_0^{\chi(z_{\rm max})} \mathrm{d}\chi \frac{W^i(\chi)W^j(\chi)}{\chi^2} P_\delta\left(\frac{\ell+0.5}{\chi}, z(\chi)\right), \tag{3.4}$$

where $P_\delta(k,z)$ is the nonlinear matter power spectrum and the lensing weight is:

$$W^i(\chi) = \frac{3H_0^2\Omega_{\rm m}}{2c^2}\frac{\chi}{a(\chi)}\int_\chi^{\chi_{\rm H}} \mathrm{d}\chi'\, n^i\left(z(\chi')\right)\frac{\mathrm{d}z}{\mathrm{d}\chi'}\frac{\chi'-\chi}{\chi'}, \tag{3.5}$$

with the source galaxy redshift distribution $n^i(z)$ normalised to integrate to 1, and $\chi_{\rm H}$ the horizon distance. We follow Krause et al. (2021), and model P_δ using a combination of CAMB (Lewis et al., 2000) for the linear part, and HaloFit (Takahashi et al., 2012) for nonlinear modifications. Even though at very small ℓ the power spectra of convergence, C_κ, and cosmological shear, C_γ, differ by a factor of $\sqrt{(\ell+2)(\ell-1)/(\ell(\ell+1))}$ (see Hu 2000 and Fosalba et al. 2015), for the scales covered in this work they are approximately the same, and can be modelled fairly simply as described in Eq. (3.4). In practice, however, $\xi_\pm$ measurements are sensitive not only to the pure cosmological shear, but also to additional correlations due to, e.g., intrinsic alignments. In the presence of IAs, the convergence spectra in Eqs. (3.2) and (3.3) are replaced by C_γ, the calculation of which we come to in Section 3.2.2.

3.2.2 Modelling Intrinsic Alignments

In general terms, the impact of intrinsic alignments (IAs) can be thought of as adding a coherent additional component to each galaxy's shape. That is, in the limit that all terms in the equation are $\ll 1$, the observed ellipticity can be written as $\gamma^{\rm obs} = \gamma^{\rm G} + \gamma^{\rm I} + \epsilon_{SN}$, or the sum of a shear due to cosmological lensing, an IA-induced shear, and a random shape noise component. Although the latter is typically dominant for any single galaxy, it cancels when the ellipticity is averaged across a large population of galaxies. At the level of angular correlation functions, one has:

$$C_\gamma^{ij}(\ell) = C_\kappa^{ij}(\ell) + C_{\rm GI}^{ij}(\ell) + C_{\rm GI}^{ji}(\ell) + C_{\rm II}^{ij}(\ell). \tag{3.6}$$

The first term, C_κ, is the auto-correlation of cosmological lensing, and is defined in Eq. (3.4). The intrinsic-intrinsic contribution is referred to as the II term, and arises from galaxies that are spatially close to one another. The intrinsic-shear cross-correlation is known as the GI term, and emerges from the fact that galaxies at different distances along the same line of sight can either be lensed by, or experience gravitational tidal interaction with, the same large scale structure.

Again assuming the Limber approximation, the angular power spectra can be written as

$$C_{\rm II}^{ij}(\ell) = \int_0^{\chi_H} \mathrm{d}\chi \frac{n^i(\chi)n^j(\chi)}{\chi^2} P_{\rm II}\left(k=\frac{\ell+0.5}{\chi}, \chi\right) \tag{3.7}$$

and

$$C_{\rm GI}^{ij}(\ell) = \int_0^{\chi_H} \mathrm{d}\chi \frac{W^i(\chi)n^j(\chi)}{\chi^2} P_{\rm GI}\left(k=\frac{\ell+0.5}{\chi}, \chi\right). \tag{3.8}$$

Given Eqs. (3.4), (3.7) and (3.8), we have the ingredients to use Eqs. (3.2) and (3.3) to predict the observable $\xi_\pm$. Note that the GI and II power spectra are model dependent. Indeed, how one calculates them is a significant analysis choice in any cosmic shear analysis. In the sections below we describe the two model choices explored in this work.

TATT Model

The tidal alignment and tidal torque model (TATT; Blazek et al. 2019) is based on nonlinear perturbation theory, which is used to expand the field of intrinsic galaxy shapes $\gamma^{\rm I}$ in terms of the tidal field s and the matter overdensity δ. Whereas δ is a scalar at all points in space, $\gamma^{\rm I}$ and s are 3×3 matrices, defining an ellipsoid in 3D space. Although in principle the expansion could be extended to any order, our implementation includes terms up to quadratic in the tidal field:

$$\gamma^{\rm I}_{ij} = C_1 s_{ij} + C_2 \sum_k s_{ik} s_{kj} + b_{\rm TA} C_1 \delta s_{ij}, \tag{3.9}$$

where C_1, C_2 and $b_{\rm TA}$ are free parameters. This leads to the power spectra:

$$P_{\rm GI} = C_1 P_\delta + b_{\rm TA} C_1 P_{0|0E} + C_2 P_{0|E2}, \tag{3.10}$$

$$P^{\rm EE}_{\rm II} = C_1^2 P_\delta + 2 b_{\rm TA} C_1^2 P_{0|0E} + b_{\rm TA}^2 C_1^2 P_{0E|0E} + C_2^2 P_{E2|E2} + 2 C_1 C_2 P_{0|E2} + 2 b_{\rm TA} C_1 C_2 P_{0E|E2}, \tag{3.11}$$

$$P^{\rm BB}_{\rm II} = b_{\rm TA}^2 C_1^2 P_{0B|0B} + C_2^2 P_{B2|B2} + 2 b_{\rm TA} C_1 C_2 P_{0B|B2}. \tag{3.12}$$

The various subscripts to the power spectra indicate correlations between different order terms in the expansion of $\gamma^{\rm I}$. They can all be calculated to one-loop order as integrals of the linear matter power spectrum over k (see Blazek et al. 2019 for the full definitions). As can be seen here, the TATT model predicts both E- and B-mode II contributions. These are propagated to separate E- and B-mode angular power spectra, which enter $\xi_\pm$ in Eqs. (3.2) and (3.3). The amplitudes are defined, by convention, as:

$$C_1(z) = -A_1 \frac{\bar{C}_1 \rho_{\rm c} \Omega_{\rm m}}{D(z)} \left(\frac{1+z}{1+z_0} \right)^{\eta_1}, \tag{3.13}$$

$$C_2(z) = 5 A_2 \frac{\bar{C}_1 \rho_{\rm c} \Omega_{\rm m}}{D^2(z)} \left(\frac{1+z}{1+z_0} \right)^{\eta_2}. \tag{3.14}$$

The pivot redshift z_0 and the constant $\bar{C}_1$ are fixed to values of $z_0 = 0.62$ and $\bar{C}_1 = 5 \times 10^{-14} M_\odot h^{-2} {\rm Mpc}^2$. Again, this is a convention, such that $C_1(z)$ and $C_2(z)$ are roughly of the order of 1 for a typical population of source galaxies. The power law term in $C_1(z)$ and $C_2(z)$ adds some flexibility to capture possible redshift evolution beyond what is already encoded in the

model. Our implementation of the TATT model then has five free parameters: $A_1, A_2, \eta_1, \eta_2, b_{\rm TA}$, which we allow to vary with wide flat priors $A_1, A_2, \eta_1, \eta_2 \in [-5, 5]$, $b_{\rm TA} \in [0, 2]$. This choice of uniformative priors is motivated by the fact that IAs are very sensitive to the properties of the galaxy population, making it very difficult to derive informative priors, and resulting in a lack of directly transferable constraints in the literature for the TATT model parameters. Although intended to match up with different alignment mechanisms, in practice A_1 and A_2 capture any correlations that scale linearly and quadratically with the tidal field. The third amplitude $b_{\rm TA}$ is designed to capture the fact that galaxies over-sample densely populated regions (i.e., one cannot sample the $\gamma^{\rm I}$ field uniformly throughout the Universe).

For this work we use the DES Y3 implementation of TATT, within CosmoSIS v1.6[1] (Zuntz et al., 2015). The power spectra in Equations (3.10)-(3.12) (with the exception of the nonlinear matter power spectrum P_δ) are generated using FAST-PT v2.1[2] (McEwen et al., 2016; Fang et al., 2017).

NLA Model

Although chronologically older and more commonly used, the nonlinear alignment model (NLA; Bridle & King 2007) is a subspace of TATT. Built on the assumption that galaxy shapes align linearly with the background tidal field, it predicts:

$$P_{\rm GI} = C_1(z) P_\delta, \quad P_{\rm II} = C_1^2(z) P_\delta, \tag{3.15}$$

with the amplitude $C_1(z)$ as defined in Eq. (3.13) in our implementation. The NLA model as implemented here differs from its predecessor, the linear alignment model (Catelan et al., 2001; Hirata & Seljak, 2004; Hirata & Seljak, 2010), by the fact that P_δ in the above equations is the full nonlinear matter power spectrum (in our case generated using HaloFit), not the linear version. Unlike the original formulation, our implementation of NLA also includes a power law redshift dependence in $C_1(z)$ to capture any additional evolution beyond the basic model (as in Eq. (3.13) above). In total, our implementation of the NLA model has two free parameters, A_1 and η_1, which we vary with the priors given in Section 3.2.2.

3.2.3 Other Nuisance Parameters & Scale Cuts

Both the TATT and NLA pipelines include free parameters for redshift error and residual shear bias. We adopt the same modelling as Krause et al. (2021), giving us one Δz and one m parameter per redshift bin, or a total of eight nuisance parameters. Note however, that these parameters are prior dominated for Y3 cosmic shear-only chains, and so add relatively little to the effective dimensionality. For details about the priors see Appendix 3.A and Table 3.A.1. We also adopt the fiducial DES Y3 cosmic shear scale cuts (see Krause et al. 2021 for an explanation of how these were derived).

[1] `https://bitbucket.org/joezuntz/cosmosis/wiki/Home`; des-y3 branch of cosmosis-standard-library
[2] `https://github.com/JoeMcEwen/FAST-PT`

3.3 Creating and analysing the cosmic shear data vector

In this section we describe how we generate mock data. This is necessary to calibrate the relation between bias in cosmological parameters and statistical metrics used for model comparison, which is central to our method for model selection. Essentially we wish to create an ensemble of data vectors that span a useful range of bias in cosmological parameters and $\Delta\chi^2_{(\mathrm{df})}$ (or R), allowing us to map out the relation between the two. Our analysis framework and mock data follow the DES Y3 choices, and we use DES Y1 (the precursor data set to Y3) to define plausible IA scenarios. In Section 3.3.1 we focus on how to simulate the cosmological lensing terms (which depend on cosmology, not the IA model). Then in Section 3.3.2 we describe the IA terms (IA model-dependent). We explain how we incorporate noise into our analyses, and why it is necessary, in Section 3.3.3. Finally, we describe our sampler choices in Section 3.3.4, and in particular an approximation using importance sampling that we use to accelerate the analysis of the noisy data vectors.

3.3.1 Generating Mock Data

For a given set of input parameters, we generate a noiseless DES Y3-like cosmic shear data vector, $\boldsymbol{D}$, using the theory pipeline described in Section 3.2. We assume the fiducial Y3 redshift distributions, as presented in Myles, Alarcon et al. (2021). In all data vectors, the same input flat ΛCDM cosmology is used ($\Omega_\mathrm{m} = 0.29$, $A_\mathrm{s} = 2.38 \times 10^{-9}$, $\Omega_\mathrm{b} = 0.052$, $h = 0.75$, $n_\mathrm{s} = 0.99$, $\Omega_\nu h^2 = 0.00053$). This corresponds to $\sigma_8 = 0.79$, $S_8 = 0.77$, where $S_8 \equiv \sigma_8\sqrt{\Omega_\mathrm{m}/0.3}$. We choose these to match the marginalised mean values from the DES Y1 3×2pt chain used to generate IA samples (see Section 3.3.2 below). Note, however, that the exact values are not expected to affect our results. We also fix all the redshift and shear calibration nuisance parameters to zero when generating data vectors.

3.3.2 Choosing IA Scenarios

When constructing simulated data vectors, it is important to remember that IA model parameters are not independent. The total GI+II intrinsic alignment component in a scenario with, e.g., $A_1 = A_2 = 1$ is very different from one with $A_1 = 0.1, A_2 = 1$. As a consequence, it is possible for two sets of input IA parameters to give similar cosmological bias (when analysed with NLA), but quite different χ^2 values. Specific combinations of TATT parameter values may enhance or cancel out cosmological parameter bias, and so it is useful to sample the 5D TATT parameter space rather than scaling up individual parameters to explore the potential for cosmological parameter bias due to model insufficiency. Therefore, instead of a single mock data vector, we generate a set of 21 data vectors, all with the same cosmology, but with different possible IA scenarios. The number of mock data vectors is an analysis choice. The more we generate, the better we cover the IA parameter space, but it also increases computational costs. We verified that 21 was a sufficient number of scenarios to have samples presenting low, medium and high bias in cosmological parameters, while still being reasonable in terms of computational cost (i.e. the chains to run).

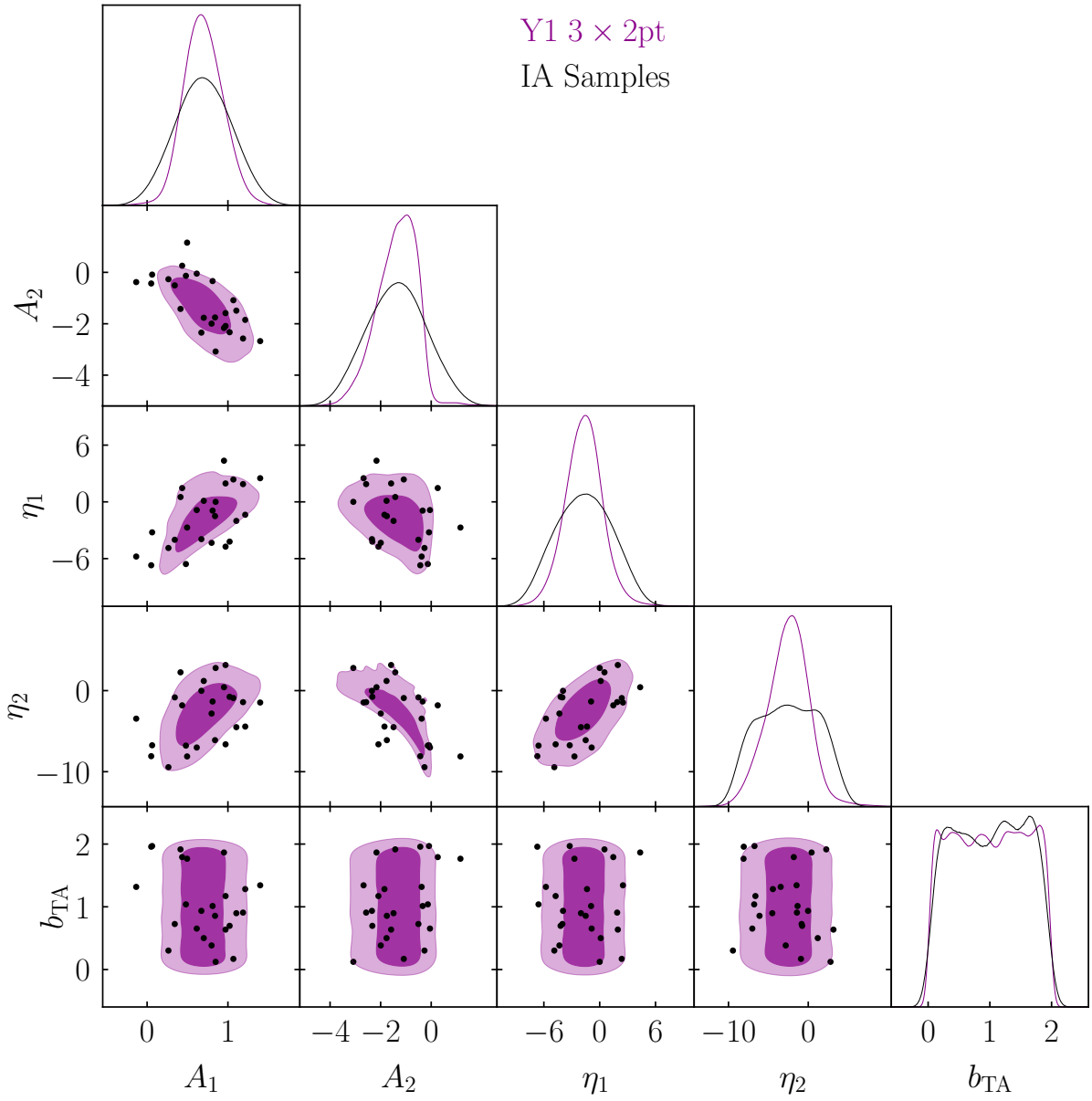

Figure 3.3.1: An illustration of how we generate samples in IA model parameter space for this work. The purple contours show the 68% and 95% confidence levels from the TATT model analysis of the DES Y1 3 × 2pt data (Samuroff et al., 2019). Overlain (black points) are the IA samples we derive from this posterior probability distribution after marginalizing over all other parameters. On the diagonal, we show the Y1 marginal posterior (purple), and also the distribution of IA samples (black), both normalised to integrate to 1 over the prior range. As shown, the latter is slightly broader than we would obtain by drawing from the DES posterior distribution.

To do this, we follow the recipe set out in Section 2 of DeRose et al. (2019). Starting with the posterior from the DES Y1 3×2pt TATT analysis (the purple contours in Figure 3.3.1; Samuroff et al. 2019), we evaluate the covariance matrix of the $N_p = 5$ TATT parameters, and perform an eigenvalue decomposition. We then use Latin Hypercube sampling to generate $N_{\rm samp}$ samples, which are roughly evenly distributed in N_p dimensional space. Finally, we use the eigenvalues/vectors to rotate and normalise those samples into the parameter space. The results are shown in Figure 3.3.1. The idea is that this provides a slightly broader coverage than could be obtained simply by drawing points from the joint posterior, while maintaining the correlations between parameters. In this way, we can cover a range of marginal cases, which are pessimistic, but still consistent with the data; this is useful, since for our purposes it is more important to span the range of plausible TATT model parameters than to preserve the statistics of the Y1 posterior exactly. Notice that the reason for using the DES Y1 posterior here, instead of the Y3 for instance, is that the method we are proposing involves using a precursor data set to generate the samples and draw conclusions about the current data set being analysed. However, the important point is to sample IA parameters in a way that approximately preserves the degeneracies seen in real data, while also covering a wide enough range to allow the bias calibration.

3.3.3 Adding Noise

Since real measurements unavoidably include an (unknown) noise realisation, the calibration of the bias-metric relation is inherently a probabilistic problem (we will return to this point in Section 4.5; see the discussion there for details). For this reason, it is important that our simulations capture all sources of scatter in the data.

For each of our 21 IA scenarios, defined by a set of input TATT parameter values $\boldsymbol{\theta}_{{\rm IA},i}$, we have a set of noisy data vectors $\tilde{\boldsymbol{D}}_{i,j} = \boldsymbol{D}(\boldsymbol{\theta}_{{\rm IA},i}) + \boldsymbol{N}_j$, where noise realisation $\boldsymbol{N}_j$ is drawn from the covariance matrix, and is assumed to be independent of $\boldsymbol{\theta}_{{\rm IA},i}$. We use the final DES Y3 covariance matrix, which is analytic and includes a Gaussian shape noise and cosmic variance contribution, as well as higher order non-Gaussian and super-sample terms (Friedrich et al., 2021). In total we generate 50 noise realisations, which we apply to each data vector. This gives us a collection of 21 noiseless data vectors, and $21 \times 50 = 1050$ noisy ones.

For testing, however, it is convenient to arbitrarily choose a single fixed noise realisation, which we refer to as *fiducial noise*. Figure 3.3.2 shows an example DES Y3-like data vector, generated using the setup described above, with the fiducial noise realisation added. For this particular example, the input TATT parameters are the mean values from the Y1 posteriors in Figure 3.3.1. For reference, we also show the noiseless version (purple solid), as well as the separate (again, noiseless) IA contributions. Since the IA signal in the lowest bin (1, 1) seems to dominate, one could reasonably ask whether we could simply focus on this part of the data vector for model selection. We ultimately decide against this for a few reasons. First, although the IA signal is strongest in bin 1, 1, we can see there is also non-negligible signal in the surrounding bins (e.g. 2, 1 and 3, 1). Indeed, during the DES Y3 analysis, a range of IA mitigation techniques were explored, including dropping the lowest auto bin correlations. For moderate TATT scenarios, it was found that this could not reliably eliminate cosmology biases, suggesting the IA contamination

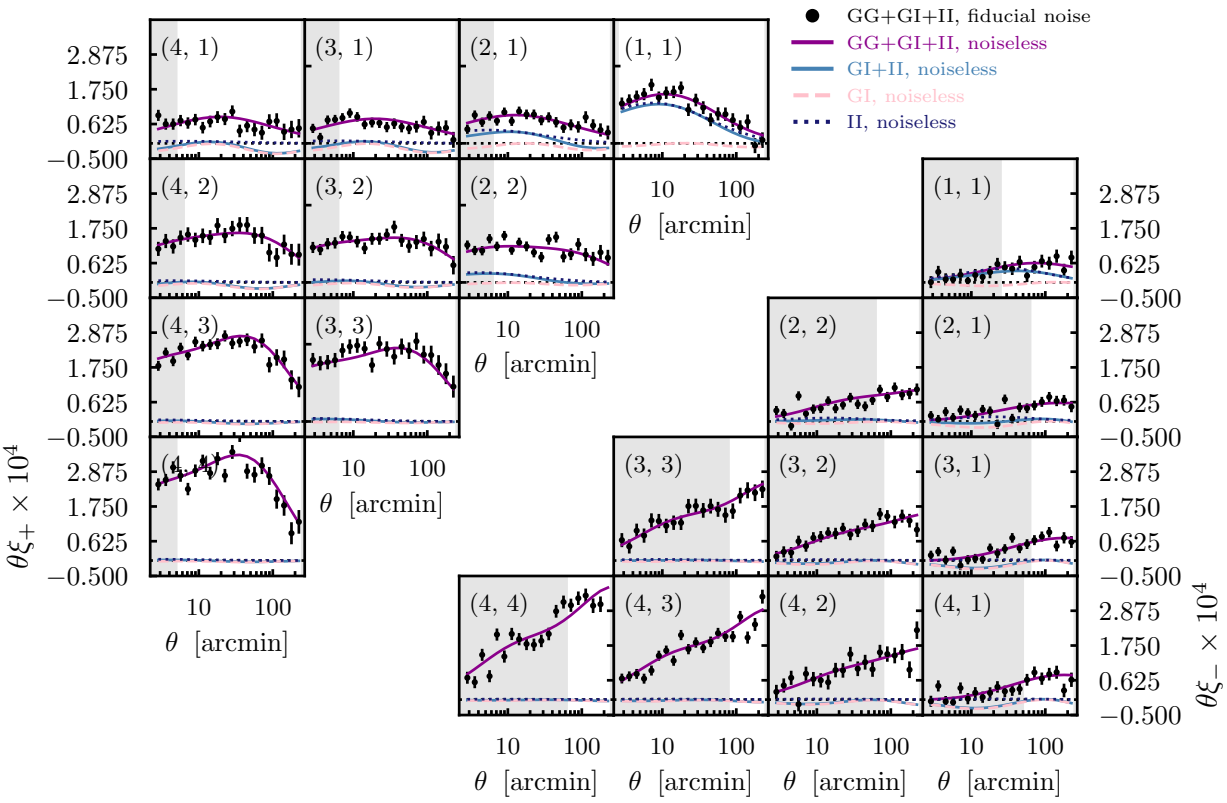

Figure 3.3.2: An example of a noisy simulated data vector of the type used in this paper. Each panel shows a redshift bin combination (as labelled), and the upper and lower triangles show ξ_+ and ξ_- respectively. In each panel we show the simulated cosmic shear data vector with fiducial noise (black points with error bars), as well as the noiseless version (smooth purple). We also show the GI and II intrinsic alignment components separately. For reference, the input IA parameters here correspond to the mean of the DES Y1 posterior discussed in Section 3.3.2 ($A_1 = 0.7, A_2 = -1.36, \eta_1 = -1.7, \eta_2 = -2.5, b_{\rm TA} = 1$). The grey bands represent the fiducial DES Y3 cosmic shear scale cuts, i.e., the scales removed from our analysis.

is not confined to the 1, 1 data vector. As we can see from Figure 3.3.2, there is also a significant GI signal in some of the more separated bin pairs (e.g. 4,1), which can dominate in some IA scenarios. Additionally, the degree to which the low redshift II contribution dominates the IA signal is somewhat dependent on the input TATT parameters, and so the strength of this assumption varies in parameter space.

3.3.4 Choice of Sampler

Nested Sampling

To sample the cosmic shear likelihood, we use the PolyChord nested sampling algorithm (Handley et al., 2015a,b), which generates estimates for the multidimensional posterior $\mathcal{P}(\boldsymbol{\theta}|\boldsymbol{D}, M)$ and the evidence for a given model $\mathcal{Z}(\boldsymbol{D}|M)$ simultaneously. This matches the DES Y3 choice, and has been validated in terms of both evidence and the contour size compared with a long Monte Carlo chain (Lemos, Weaverdyck et al., 2022). We briefly explored the possibility of using MultiNest (Feroz et al., 2019), which is conceptually similar, but significantly faster. Ultimately, however, we found that MultiNest underestimates the width of the posteriors in all cases we tested (both NLA and TATT; see Appendix 3.D). It also gives inaccurate evidence values (Lemos, Weaverdyck et al., 2022), which tend to skew towards preferring NLA. For these reasons, we did not pursue this.

To obtain estimates for the best χ^2, we use oversampled chains (i.e., output with 10× the number of points as saved in the standard chains). This approach has been tested in the Y3 cosmic shear setup, and shown to give comparable results to running a likelihood maximiser (Secco, Samuroff et al., 2022). All sampling, as well as the modelling steps described in Section 3.2, are carried out using CosmoSIS.

Importance Sampling

To assess the impact of data vector noise, in addition to nested sampling we also make use of importance sampling (IS; Neal 1998; Tokdar & Kass 2010 see also Lewis & Bridle 2002; Padilla et al. 2019 for cosmology-specific applications). For each IA scenario $\boldsymbol{\theta}_{\mathrm{IA},i}$, we wish to estimate the shape and position of the $S_8 - \Omega_\mathrm{m}$ posterior, as well as the best fit and evidence. Running full chains for every combination of noise and IA scenario would be expensive, and IS provides a fast approximation.

Say one wants to estimate the characteristics of a distribution P, over parameter θ. One can estimate the mean of the function $f(\theta)$ under P as:

$$\hat{f} = \int f(\theta) P(\theta) \mathrm{d}\theta. \tag{3.16}$$

This can be rewritten in terms of a second distribution P_0:

$$\hat{f} = \int f(\theta) \frac{P(\theta)}{P_0(\theta)} P_0(\theta) \mathrm{d}\theta \tag{3.17}$$

$$\approx \sum_{\theta_i \sim P_0} f(\theta_i) w(\theta_i), \tag{3.18}$$

where we have redefined the ratio of distributions as a weight $w = P/P_0$. The second line follows as a Monte Carlo estimate for the first, and the sum runs over values of θ drawn from $P_0(\theta)$. The equations above make no assumptions about Gaussianity, or about the nature of the distributions.

To work well, however, it does require P_0 to be non-zero over the range of θ for which we wish to estimate P, and it works better in cases where the number of samples is large. Functionally, it also requires (a) that one has, or can generate, samples from P_0, and (b) that for any given θ, one can evaluate both P and P_0.

For our application, $P_0 = \mathcal{P}(\boldsymbol{\theta}|\boldsymbol{D}, M)$ is a reference posterior obtained from running a chain on the noiseless data vector $\boldsymbol{D}(\boldsymbol{\theta}_{\rm IA})$. As before, we use the $\times 10$ oversampled PolyChord output for this. The target distribution $P = \mathcal{P}(\boldsymbol{\theta}|\tilde{\boldsymbol{D}}, M)$ is the posterior we are trying to estimate, conditioned on a noisy data vector $\tilde{\boldsymbol{D}}$. With this setup, we can estimate P for each noise realisation by simply iterating through the samples from P_0 and assigning each a weight equal to the ratio of the two posteriors.

In addition to the target posterior for each model, we also estimate the best χ^2. For this, we create a high density pool of samples by merging all of our oversampled PolyChord chains (21 IA samples), in addition to a small number of additional chains run with a Y1 like covariance matrix. This gives us over a million points in parameter space per model. For each noisy data vector, we re-evaluate the likelihood at each point, and select the maximum. Given an estimate for the best χ^2 from IS, and assuming a Gaussian likelihood, the Bayesian evidence can then be estimated as (see Section 3 Joachimi et al. 2021b):

$$\ln\mathcal{Z}(\tilde{\boldsymbol{D}}_i|M) = \ln\mathcal{Z}_0(\boldsymbol{D}|M) - 0.5\left(\chi_i^2 - \chi_0^2\right), \tag{3.19}$$

where index i indicates a noise realisation, and $\mathcal{Z}_0$ and χ_0^2 are the evidence and best χ^2 obtained from a fiducial reference chain, which in our case are our noiseless chains.

To test the performance of our IS setup, we ran five additional PolyChord chains at different noise realisations, once in the low bias regime and again in the intermediate ($\sim 1\sigma$) bias regime. We verified that in all cases, our IS setup recovered the best χ^2 as well as the shape and mean of the posteriors with a comparable level precision to a full chain. Our implementation is a slight modification of the code discussed in Weaverdyck, Alves et al. (2022), and will be available on release of that paper.

3.4 Model Selection

In this section, we define the components of our model selection method. In essence, we are proposing to calibrate the observed value of model comparison statistics against the probability of cosmological parameter bias. These quantities can be computed using noisy simulated data, but first they must be properly defined. To this end, we define how we quantify cosmological parameter bias in Section 3.4.1. The metrics that we tested in search of a useful bias–metric relation are discussed in Section 3.4.2. In Section 3.4.3 we make considerations regarding unconverged samples. We summarise our model selection method in Section 3.4.4.

3.4.1 Significance Level of Cosmological Parameter Biases

Now that we have a set of noisy data vectors, the next step is to fit them with all parameters free. We analyse each data vector twice, fitting to our full set of cosmological and nuisance parameters, but in one case using TATT, and in the other NLA. We then define bias as the distance between the peak of the marginalised posteriors in the $S_8 - \Omega_m$ plane. Figure 3.4.1 illustrates this for a particular simulated data vector. In brief, the algorithm works by evaluating the Euclidean distance between the peaks of the two posteriors, $\mathcal{P}_1$ and $\mathcal{P}_2$ in $S_8 - \Omega_m$ space. It then sequentially computes the confidence ellipses of $\mathcal{P}_1$ at different σ levels, and finds a value of N_σ that minimises the distance between the ellipse and the peak of $\mathcal{P}_2$. Note that this is the same recipe used in Krause et al. (2021). The choice of S_8 and Ω_m as the parameters of interest comes simply from the fact that these are the cosmological parameters best constrained by DES. One could conceivably use a more complicated separation metric that is sensitive to the full parameter space, along the lines of those used for assessing tensions between data sets (e.g. Lemos, Raveri, Campos et al. 2021). Conversely, given limitations in constraining power and the potential impact of the priors on some parameters (e.g. Ω_m), one could opt to use the 1D offset in the best-constrained parameter. For cosmic shear this would be S_8, but this choice would vary depending on the modelling choices, the nature of the data being analysed, and the overall scientific goal of the analysis. For our purposes, however, we follow Krause et al. (2021) and consider the simpler 2D metric to be sufficient.

We use noisy simulated data, as described in Section 3.3.3 – this is an important feature of our analysis, and it is necessary to allow us to meaningfully interpret our statistical metrics. Therefore, the relative separation of the two posteriors is a more useful quantity than the distance from the input values of S_8 and Ω_m. The 0.22σ value shown in Figure 3.4.1 is assessed relative to the TATT posterior. This found to be more stable than assessing it relative to the NLA posterior, particularly in relatively extreme IA scenarios where the NLA posteriors are significantly shifted and can be distorted by prior edge effects.

Finally, it is implicit in the above that marginalised TATT constraints represent correct results by which to measure bias. That is, when we refer to *bias*, we are in fact talking about *bias in the cosmological model when assuming the NLA model, with respect to what we find when assuming the TATT model*. Although this is clearly reasonable (since our data were created using TATT), marginal contours can be subject to projection effects. Indeed, since some of the TATT parameters are relatively poorly constrained by shear-shear analysis alone, the two IA models cannot be assumed to experience projection effects to the same degree. We test this in Appendix 3.A, and find projection offsets between TATT and NLA at the level of 0.1σ. This is well below the threshold of 0.3σ used for this work, and is thus unlikely to significantly affect our results.

3.4.2 Model Comparison Statistics

We investigate two commonly used test statistics, the χ^2 difference and the Bayes ratio. We show in Section 4.5 that the former is more robust against noise than the latter, and is therefore a more useful metric for the method we are proposing.

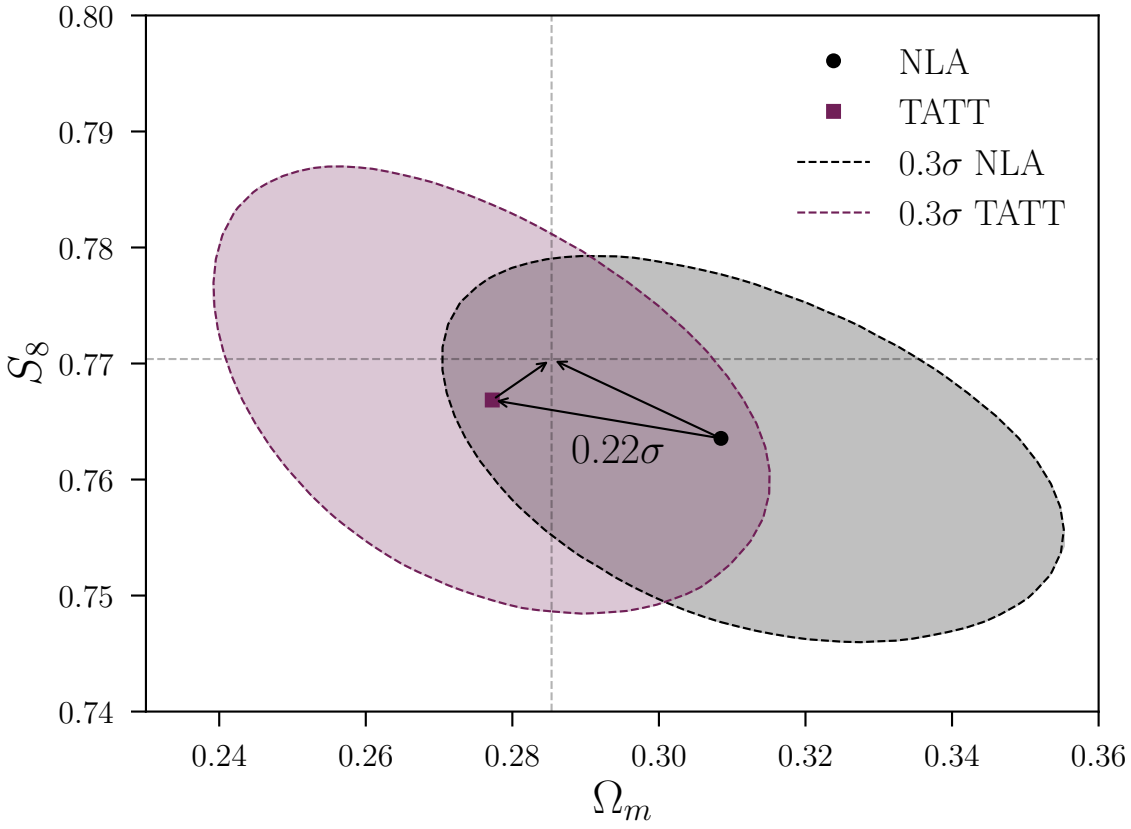

Figure 3.4.1: An example of how cosmological parameter bias is defined for a given IA scenario and noisy data vector. The purple point and the dotted ellipsoid show the maximum likelihood and 0.3σ contour, obtained from the analysis of a noisy simulated data vector with the TATT model. The black is the same, but with the NLA model. The vector connecting the two peaks in the $\Omega_{\rm m} - S_8$ plane defines our bias metric. Note that the TATT contour is slightly offset from the input parameter values (the dashed lines) due to noise and projection effects. It is for this reason that the relative separation, rather than the distance from the input, is the most appropriate bias definition.

χ^2 Difference Tests

When dealing with nested models (i.e., where one model is a subspace of the other, as in the case with NLA and TATT), the difference in the best χ^2 that can be achieved by each model on the same data is a convenient statistic for model selection (Steiger et al., 1985; Rigdon, 1999; Schermelleh-Engel et al., 2003; Andrae et al., 2010).

The χ^2 difference metric is defined as the difference in the best χ^2 values of the parameter fits when assuming the two IA models, divided by the difference in their numbers of degrees of freedom (df):

$$\Delta\chi^2_{(\mathrm{df})} = \frac{\chi^2_s - \chi^2_l}{df_s - df_l}, \tag{3.20}$$

where s indicates the *smaller* model (the one with fewer free parameters and therefore more degrees of freedom; NLA in our example), and l denotes the model with more parameters, *larger*, and so fewer degrees of freedom (TATT in our case)[3]. Note that the point estimate to use for the "best χ^2" here is somewhat subjective. For a given chain, we use the value closest to the peak of the multidimensional posterior. One could also use the maximum likelihood, although in practice this tends to be a slightly noisier quantity.

In the limit that the additional parameters in the larger model have no impact on the quality of the fit, the metric is exactly zero: $\Delta\chi^2 = \chi^2_s - \chi^2_l = 0$. Very small $\Delta\chi^2_{(\mathrm{df})}$ values can therefore be taken as evidence that the smaller model is sufficient, given the data. In practice, however, this is an unlikely outcome, as extra parameters will typically allow the model more flexibility. Under the null hypothesis that the two models s and l both adequately fit the data, the value of the numerator $\chi^2_s - \chi^2_l$ is χ^2-distributed with $df_{\mathrm{diff}} = df_s - df_l$ degrees of freedom, and the expectation value is $\langle\Delta\chi^2\rangle \sim 1$ (Wilks, 1938). One can interpret larger $\Delta\chi^2$ using the corresponding *p-values* to quantify the degree to which the data appear to favour the larger model. As we will discuss in Section 4.5, however, for our purposes it is more useful to focus on the observed relation between $\Delta\chi^2_{(\mathrm{df})}$ and parameter bias than on formal statistical thresholds. That is, we propose to use $\Delta\chi^2_{(\mathrm{df})}$ as an empirical tool, which requires calibration using simulated likelihood analyses for any given problem. This way, we are also free from other assumptions behind the standard use of the $\Delta\chi^2_{(\mathrm{df})}$ metric – for example, it formally requires nested models whereas an empirical calibration would not. Note that our approach here is functionally similar, but motivated slightly differently, to the calibration of Posterior Predictive Distribution (PPD) $p-$values for internal consistency testing, as implemented in Doux, Baxter et al. (2021).

In principle, $\Delta\chi^2_{(\mathrm{df})}$ is prior-independent. In Bayesian inference, however, the prior typically controls the regions of parameter space that can be explored, and so restrict the values of χ^2 that can be attained. In practice, this is only an issue if the likelihood peaks outside the prior bounds (which is, in any case, usually a red flag).

[3]Note that it is this quantity, weighted by the difference in df, that we refer to as $\Delta\chi^2_{(\mathrm{df})}$ throughout the paper, and not the simple $\chi^2_s - \chi^2_l$ difference. This allows us to briefly compare with theoretical cut-offs in Section 3.5.1. For practical purposes, however, this is not strictly necessary – one could just as easily calibrate the raw difference.

One other point to remember is that, although in an ideal case with well-constrained parameters one extra parameter constitutes one fewer degree of freedom, in practice this is often not true. In such cases, one can calculate the *effective* degrees of freedom (see Raveri & Hu 2019). With the fiducial DES Y3 cosmic shear setup (minus the shear ratios), the effective degrees of freedom for TATT and NLA are 222 and 224 respectively, giving $df_{\rm diff} = 2$ (compared with $df_{\rm diff} = 3$ from simple parameter counting; see Secco, Samuroff et al. 2022 Table III).

In Section 4.5, we also briefly consider two other likelihood-based metrics: the Akaike Information Criterion (AIC; Akaike 1973) and the Bayesian Information Criterion (BIC; Schwarz 1978). Although these statistics have very different theoretical underpinnings (see Liddle 2007), they are similar in form, and can be conveniently reformulated as thresholds in $\Delta\chi^2_{\rm (df)}$. As with $p-$value cut-offs, however, they are seen to be relatively under-cautious in separating high- and low-bias scenarios (see Section 4.5 and Figure 3.5.1).

Bayes Ratio

The Bayesian evidence ratio, or Bayes ratio R (Jeffreys, 1961; Kass & Raftery, 1995), is a slightly more complicated alternative to $\Delta\chi^2_{\rm (df)}$. It is defined as the probability of measuring a data vector $\boldsymbol{D}$ assuming a given model M_1, divided by the probability of measuring the same data $\boldsymbol{D}$ for a second model M_2:

$$R \equiv \frac{\mathcal{Z}(\boldsymbol{D}|M_1)}{\mathcal{Z}(\boldsymbol{D}|M_2)}. \tag{3.21}$$

Here, $\mathcal{Z}(\boldsymbol{D}|M)$ is the Bayesian evidence, which can be obtained marginalising over all the model parameters $\boldsymbol{\theta}$:

$$\mathcal{Z}(\boldsymbol{D}|M) = \int \mathrm{d}\boldsymbol{\theta}\ \mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M)P(\boldsymbol{\theta}|M), \tag{3.22}$$

where $\mathcal{L}(\boldsymbol{D}|\boldsymbol{\theta}, M)$ is the likelihood, and $P(\boldsymbol{\theta}|M)$ is the prior, both assuming a particular model. The Bayes ratio is typically interpreted using the Jeffreys scale (Jeffreys, 1961), which defines ranges of values that match up to labels (e.g., “strong evidence", “substantial evidence", etc.).

Note that R and $\Delta\chi^2_{\rm (df)}$ are not independent from one another (indeed the latter approximates the former under certain assumptions; see Bishop 2006; Marshall et al. 2006). It is important, then, to be careful when seeking to combine information from the two.

Evidence ratios have been widely used in cosmology, both for comparing different data sets under the same model (i.e., as a tension metric; Marshall et al. 2006; Lemos, Raveri, Campos et al. 2021), and for model comparison on the same data (Liddle, Mukherjee & Parkinson, 2006a; Kilbinger et al.,, 2010; Secco, Samuroff et al., 2022). It is worth bearing in mind that the formulation in the two contexts is slightly different. In the former case there is explicit prior dependence, which motivates the use of statistics such as Suspiciousness (see e.g. Lemos, Raveri, Campos et al. 2021 Section 4.2). The version commonly used for model selection, on the other hand, should be independent of the choice of priors, at least in the limit that (a) the models are nested and (b) the priors on the extra parameters are wide compared with the likelihood.

Since cosmological analyses involve a large number of free parameters, computing the Bayesian evidence requires integrating a probability distribution over a high number of dimen-

sions. A common way to calculate it is while producing the posterior distributions, using nested sampling (Skilling, 2006). The precision required from the sampler in order to compute reliable Bayesian evidences, however, often makes the sampling time very long. We choose to use the PolyChord nested sampling algorithm in this work (Handley et al., 2015a,b) – although see Appendix 3.D, where we consider the feasibility of using MultiNest (Feroz et al., 2019) as a slightly faster alternative.

Bias Probability

The above quantities give us the basic tools for our model comparison. There is, however, a piece missing. As we mention in Section 3.3.3, the calibration is inherently probabilistic. The model comparison metrics (both $\Delta\chi^2_{(\mathrm{df})}$ and R), as well as the offset between the NLA and TATT best fits, are somewhat sensitive to noise, and we do not know the true noise realisation in the data. We thus define a *bias probability* P for a particular bias tolerance X:

$$P(b > X\sigma|\Delta\chi^2_{(\mathrm{df}),\mathrm{obs}} < \Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}) = \frac{N^{b>X}_{\mathrm{samp}}}{N^{b>X}_{\mathrm{samp}} + N^{b<X}_{\mathrm{samp}}}. \tag{3.23}$$

In words, P is the probability of the bias in S_8–Ω_m being greater than $X\sigma$, *if* the observed $\Delta\chi^2_{(\mathrm{df})}$ from the data is below some threshold $\chi^2_{(\mathrm{df}),\mathrm{thr}}$ (which is to be determined empirically based on the adopted X and P).

It is estimated by plotting the distribution of all of our noisy data vectors in the bias–$\Delta\chi^2_{(\mathrm{df})}$ plane, and, for a particular $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$, evaluating the fraction of points that lie both *above* bias = $X\sigma$ and *below* $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$ (i.e., in the lower right quadrant of Figure 3.5.1). In practice, one starts by defining the tolerance X and the desired bias confidence P. For example, one might require the bias to be smaller than $X = 0.3\sigma$ at 90% confidence. Given those numbers, we can then iteratively evaluate Eq. (3.23) with different $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$ thresholds until the required P is achieved.

3.4.3 Dealing with Unphysical $\Delta\chi^2_{(\mathrm{df})}$ Values and Unconverged IA Samples

It is also worth briefly remarking that in our analysis we found about 50 (out of 1050) data points for which $\chi^2_{\mathrm{TATT}} > \chi^2_{\mathrm{NLA}}$, and so $\Delta\chi^2_{(\mathrm{df})} < 0$. Given that NLA and TATT are nested models, these points are unphysical (a more flexible model should always be able to produce a better or as-good fit). We conclude that they are an artefact of the sampling method; although we tested the robustness of our IS setup, and found it can reproduce the best fit from chains to reasonable precision, some level of sampling noise is still present. Given this logic, it is reasonable to assume that if we were to find $\Delta\chi^2_{\mathrm{NLA}} < \chi^2_{\mathrm{TATT}}$ in real data, there would likely be some follow-up investigation and the chains would be rerun. This is particularly true if the χ^2 is an integral part of the analysis plan, as it is in our method. We thus choose to discard these points. It is worth bearing in mind, however, that this may or may not be a reasonable decision in other setups, depending on the models being compared and the details of the analysis.

Also note that, although our results are based on 21 IA samples, we did initially draw 25 scenarios (see Section 3.3.2). Of these 25, we found four to be so extreme that the NLA PolyChord chains failed to converge in the noiseless case. These resulted in highly distorted and often bimodal contours in the $S_8 - \Omega_m$ plane, making it difficult to obtain meaningful estimates for the bias. Given this, and also for the reasoning discussed above, we choose to omit these samples from our analysis. This leaves us with a total of 21 IA samples.

3.4.4 The Recommended Method for Model Selection

Given all the definitions set out in the sections above, we now follow the recipe outlined below, in order to map and calibrate the bias-metric relation. These steps are, in essence, our method; when written out in this form, it can be very easily generalised to other model selection problems beyond our particular example of IA in cosmic shear.

1. **Sample IA scenarios:** Draw about $10 - 30$ parameter samples from either a posterior from a previous analysis or from some reasonable priors using the method described in Section 3.3.2 (we used 21 drawn from DES Y1 TATT posteriors).

2. **Generate data vectors:** Generate a simulated noiseless data vector for each IA scenario drawn in the previous step. Other model parameters (e.g., cosmological and nuisance parameters) should be fixed to some fiducial values. See Section 3.3.1.

3. **Analyse noiseless data vectors:** There are two chains per data vector, one corresponding to model M_1, and the other to M_2. Again, we compared the TATT and NLA IA models using PolyChord to compute statistics. These choices might vary under different applications. Details on the sampling can be found in Section 3.3.4.

4. **Compute parameter bias and plot out the bias-metric relation:** Demonstrate that the noiseless data vectors show a clear correlation between the test statistic (e.g., $\Delta\chi^2_{(\mathrm{df})}$ or Bayes factor) and parameter bias, as described in Section 3.4.1.

5. **Generate noise realisations:** For each data vector, generate ~ 50 noise realisations using the covariance matrix, as explained in Section 3.3.3. In our case, that gives a total of $21 \times 50 = 1050$ noisy data vectors.

6. **Analyse noisy data vectors:** As discussed above, we choose to use importance sampling to give a fast approximation for the noisy posteriors. For all noise realisations (50) and IA scenarios (21), for both IA models (2) $\longrightarrow$ 2100 total, estimate the posterior, the NLA-TATT bias and the model test statistics. See Section 3.3.4.

7. **Calculate probability:** Plot out the bias-metric relation. Use the quantities computed in the previous step to calculate the probability of bias greater than some pre-defined threshold $X\sigma$ (see Section 3.4.2).

8. **Run analyses on blinded data:** Run a full chain on the real data in order to obtain the observed model statistic.

9. **Select the model:** Interpret the observed model statistic from the previous step in terms of bias probability using the results of step vii. If the probability of exceeding the chosen level of cosmological parameter bias is low enough (where both the level of cosmological bias, and the bias probability, are analysis choices), one can safely opt for the simpler model. See Section 4.5.

Note that, to obtain an accurate calibration of the $\Delta\chi^2_{(\mathrm{df})}$ value, all aspects of the modelling should be as close to the final fiducial analysis setup as possible. For an estimate of the computational resources required to employ the proposed method, see Appendix 3.E.

3.5 Results

Now we have outlined the details of our method in Section 3.4.4, we will now consider a specific application. As discussed earlier, we choose to focus on the problem of deciding between two intrinsic alignment models for a cosmic shear analysis: NLA and TATT. Although these models *are* nested, the method does not assume this. Indeed, the only requirement is the ability to generate mock data to calibrate the chosen test statistics; therefore it is quite general and can be applied to a variety of model selection problems.

In Section 3.5.1 we discuss the results from our PolyChord chains on noiseless data vectors, and the basic trends. Section 3.5.2 then discusses the more complete probabilistic calibration, which properly factors in the impact of noise. We also compare our empirical results against theoretically derived χ^2 thresholds. Section 3.5.3 looks at how far bias can be inferred from NLA fits alone, without explicit model comparison. Lastly, Section 3.5.4 considers the wider outlook for lensing cosmology.

3.5.1 The Noiseless Case

Considering first the noiseless case, Figure 3.5.1 shows the relation between bias in the $S_8 - \Omega_\mathrm{m}$ plane and the NLA-TATT $\Delta\chi^2_{(\mathrm{df})}$. Each point results from running two chains on the same noiseless simulated data vector, first using NLA, and then using TATT. As defined in Eq. (3.20), large values of $\Delta\chi^2_{(\mathrm{df})}$ indicate statistical preference for the larger model (i.e., TATT). We see a relatively tight relation between bias and $\Delta\chi^2_{(\mathrm{df})}$, going from $\Delta\chi^2_{(\mathrm{df})} < 1$ when bias is small to relatively large values at the high bias end: low bias $\rightarrow$ small χ^2 difference, high bias $\rightarrow$ large χ^2 difference. Interestingly, the relation appears to have the form (approximately) of a double power law, with a steep slope in the high bias regime, switching to a somewhat shallower function below 0.3σ. It is worth stressing, however, that this relation is empirical. We do not have a particular expectation for its shape, and it is likely that the details depend on the analysis choices and survey setup. Note that even without data vector noise, this relation presents some scatter. This arises both from sampler noise, and from the fact that this is a complex high dimensional problem,

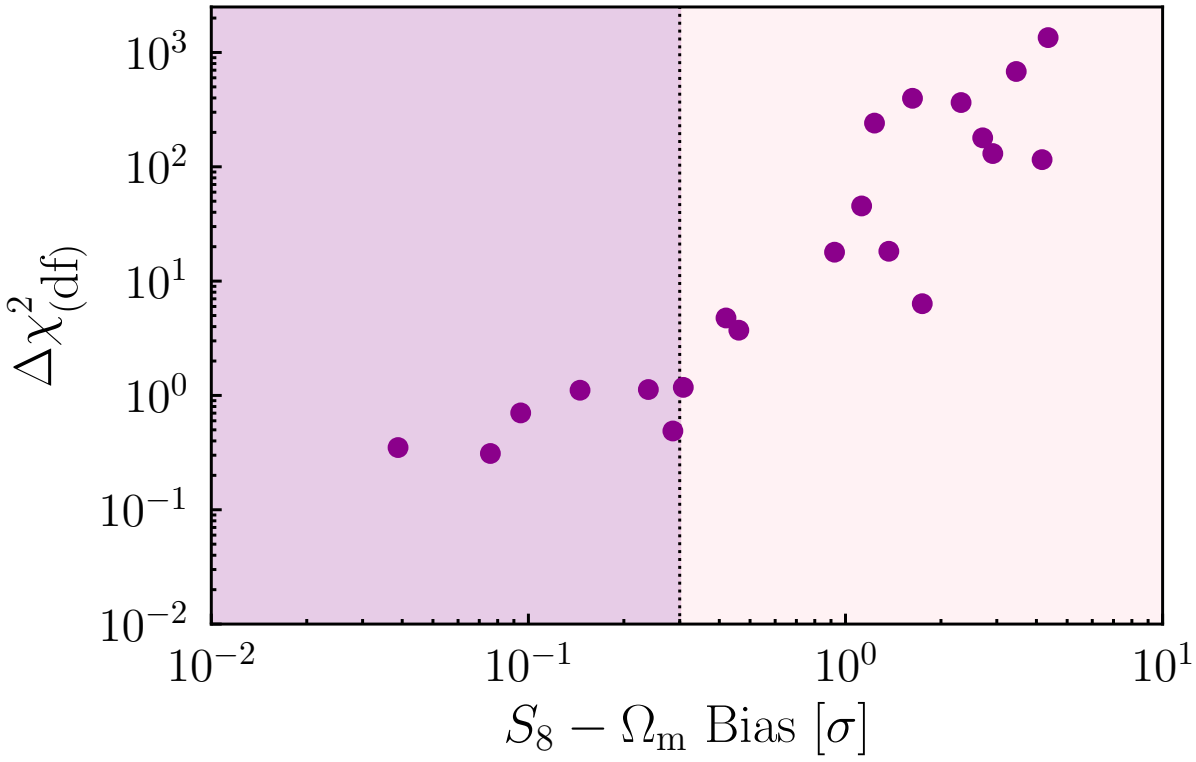

Figure 3.5.1: $\Delta\chi^2_{(\mathrm{df})}$ as a function of cosmological parameter bias for a DES Y3-like cosmic shear analysis. The 21 points correspond to noiseless data vectors, generated with different input IA parameters. As defined in Eq. (3.20), large values of $\Delta\chi^2_{(\mathrm{df})}$ indicate that the data prefer TATT over NLA. The vertical dotted line marks the 0.3σ bias limit used in DES Y3 (Krause et al., 2021). We see a clear correlation between the observable metric ($\Delta\chi^2_{(\mathrm{df})}$) and the underlying parameter bias, particularly for those points for which the bias exceeds $\sim 0.2\sigma$.

for which two sets of IA values that produce biases of a similar magnitude will not necessarily produce identical $\Delta\chi^2_{(\mathrm{df})}$ values. The vertical dotted line marks the 0.3σ bias threshold adopted by DES (Krause et al., 2021), which we adopt as our fiducial tolerance (see Section 3.5.2 below). Although we cannot use this noiseless result for any empirical method because real data will always contain noise, confirming that these quantities clearly correlate is a necessary first step in our method, and important to check before incurring the expense of further calculations. We will see in the next section that the correlation between $\Delta\chi^2$ and $S_8 - \Omega_\mathrm{m}$ bias holds (with some additional scatter) when we proceed to the noisy case.

In addition to the $\Delta\chi^2_{(\mathrm{df})}$, we also consider the Bayes Ratio as a potential model comparison metric; while the former presents a clear relation to the bias (as seen in Figure 3.5.1), we find the latter be a relatively weak indicator, with additional intrinsic scatter. This can be seen in Appendix 3.C, and in particular Figure 3.C.1. We also note that R and $\Delta\chi^2_{(\mathrm{df})}$ are correlated. In principle one could seek to combine them, but naively treating them as independent metrics is almost certainly double-counting information. Therefore, here we focus on the results using $\Delta\chi^2_{(\mathrm{df})}$. For further discussion of the Bayes ratio see Appendix 3.C.

3.5.2 Noise & Probabilistic Calibration

In Figure 3.5.2 we illustrate the impact of data vector noise in the bias-$\Delta\chi^2_{(\mathrm{df})}$ plane. We show the same 21 noiseless samples discussed above, but now overlain with multiple different noise realisations, as approximated using importance sampling. As we can see, noise introduces scatter in the bias-$\Delta\chi^2_{(\mathrm{df})}$ relation. While this noise is considerably less than in the case of the Bayes factor (for which we show in Appendix 3.C that the scatter due to noise is so large that the relation with bias is extremely weak), it is still non-negligible.

For comparison, we show the $\Delta\chi^2_{(\mathrm{df})}$ cut-offs implied by some standard model selection metrics: BIC, AIC and a $p-$value significance threshold[4] of $p = 0.05$ (see Section 3.4.2 for definitions). Unfortunately, in the presence of noise, we see that all three cut-offs are relatively weak indicators of bias – i.e., they still favour the simpler model even when significant amount of bias has been introduced in the cosmological parameters. Even in the case of AIC, which is the strictest of the three, there are a problematic fraction of noise realisations where the observable metric favours NLA, and yet NLA is biased by $> 0.3\sigma$ (see the points in the lower right hand corner of Figure 3.5.2). This illustrates a key motivation for adopting an empirical calibration. Theoretical limits imposed using, e.g., $p-$values are not designed to optimise the quantities we care most about (i.e., parameter biases). For a given analysis, it is impossible to know from first principles what level of bias is excluded for a given statistical metric cut-off without some form of calibration.

These observations have important consequences for our method. Since the exact noise realisation in any real data set is unknown, one cannot simply run a single set of IA samples (as in Figure 3.5.1), and perform a 1:1 bias-$\Delta\chi^2_{(\mathrm{df})}$ mapping. Nor, as we can see from Figure 3.5.2,

[4]Assuming that the TATT model has 2 additional effective degrees of freedom compared with NLA. Note that this df_diff value was calculated for the DES Y3 shear-only (no shear ratio) case in Secco, Samuroff et al. (2022). It is thus valid for our particular case, but would not necessarily hold under changes to the data vector or analysis choices.

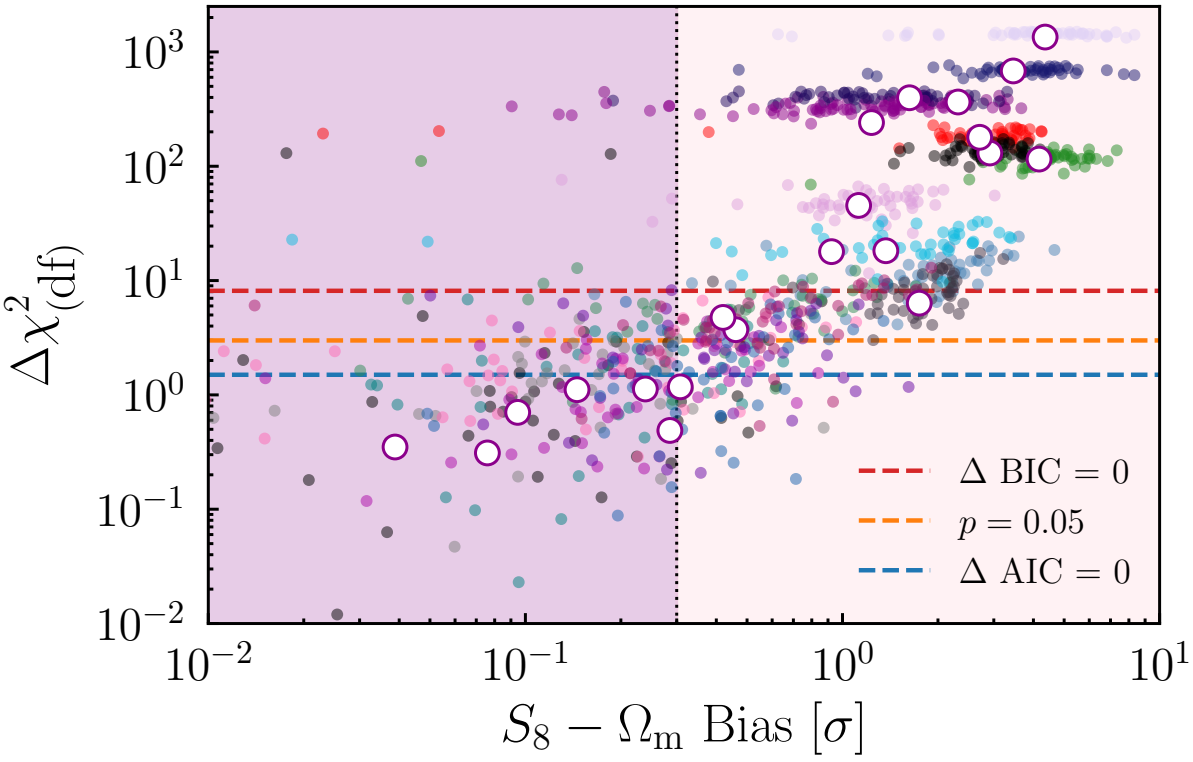

Figure 3.5.2: The impact of data vector noise on $\Delta\chi^2_{(\mathrm{df})}$. The larger open points show our 21 IA samples with zero noise (identical to those in Figure 3.5.1). The smaller coloured dots show the effect of adding random noise realisations, for which parameter constraints are estimated using importance sampling. For each of the 21 colours, we have a collection of 50 realisations. The red and blue horizontal dashed lines mark threshold $\Delta\chi^2_{(\mathrm{df})}$ values, defined by the points where the BIC and AIC respectively prefer NLA and TATT equally. The orange dashed line corresponds to a $p-$value $p(\Delta\chi^2) = 0.05$ (see text, Section 3.5.2). The fact that these formal cut-offs are relatively ineffective in isolating the bias $< 0.3\sigma$ region motivates us to adopt an empirical approach.

can we simply fall back on theoretical cut-offs to reliably guard against model bias. Instead, we must consider the problem as a probabilistic one, and factor in the uncertainty from noise.

Probabilistic Interpretation

To interpret our results in a quantitative way, we use Eq. (3.23) and calculate the bias probability $P(b > X\sigma|\Delta\chi^2_{(\mathrm{df}),\mathrm{obs}} < \Delta\chi^2_{(\mathrm{df}),\mathrm{thr}})$. This quantity should be interpreted as the conditional likelihood that, *if* in the real data one finds a $\Delta\chi^2_{(\mathrm{df}),\mathrm{obs}}$ value below some limit $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$ (a horizontal line in Figure 3.5.2), then the analysis using NLA will still in fact be biased by $X\sigma$ or more.

Figure 3.5.3 shows three curves corresponding to chosen bias thresholds of 0.15σ, 0.3σ, and 0.5σ. Each point is calculated using Eq. (3.23) and the curves are obtained by fitting a fifth order polynomial to the points. We tested the stability of these smoothed fits, and found that they are robust to doubling the number of noise realisations in Figure 3.5.2 (from 50 per IA sample to 100). This result provides a powerful tool, which can be used to interpret results from real data. For instance, say we were to run NLA and TATT chains on a blinded Y3 data vector, and find $\Delta\chi^2_{(\mathrm{df}),\mathrm{obs}} < 0.4$. With the aid of Figure 3.5.3, we could say that the chance of the NLA run being biased by more than 0.5σ in $S_8 - \Omega_\mathrm{m}$ is about 3%. The probability of exceeding a 0.3σ threshold is about 10%, and the chance of bias greater than 0.15σ is about 37%. In practice, the bias tolerance is an analysis choice. As discussed previously, DES Y3 chose a value of 0.3σ by which to judge simulated chains. The exact number, however, is somewhat subjective, and the most convenient value may depend on how well sampled the low bias end of the bias-$\Delta\chi^2_{(\mathrm{df})}$ relation is. As one might expect, the lower the bias threshold X, the stronger the requirement on the $\Delta\chi^2_{(\mathrm{df})}$ (i.e., the stronger the data needs to favour NLA) in order to keep the bias probability $P(b > X\sigma)$ low.

To understand how our results depend on various analysis choices, it is perhaps useful to think of the process in Section 3.4.4 as a series of transformations between different distributions. The points in Figure 3.5.2, which determine the final χ^2 threshold, can be thought of as the convolution of two parts: an initial distribution of noiseless points $P_s(\Delta\chi^2_{(\mathrm{df})}, b)$ (the open points in Figure 3.5.2 and the filled in Figure 3.5.1) and a second distribution conditioned on each one $P_N(\widetilde{\Delta\chi^2}_{(\mathrm{df})}, \tilde{b}|\Delta\chi^2_{(\mathrm{df})}, b)$ (where the tilde denotes noisy values of $\Delta\chi^2_{(\mathrm{df})}$ and bias). In the first case, P_s, we start with a distribution in IA parameter space $P(\boldsymbol{\theta}_\mathrm{IA})$, which we choose. The samples from $P(\boldsymbol{\theta}_\mathrm{IA})$ are mapped onto a distribution of noiseless data vectors, which are then transformed (via running chains) into samples in the final bias-$\Delta\chi^2_{(\mathrm{df})}$ space: $P(\boldsymbol{\theta}_\mathrm{IA}) \rightarrow P(\boldsymbol{D}) \rightarrow P_s(\Delta\chi^2_{(\mathrm{df})}, b)$. Both mapping steps are dependent on the survey analysis choices (choice of power spectrum, $n(z)$, covariance matrix, etc.). This is not a problem, as long as these choices match the ones that will be applied on real data. It is, however, likely these choices have an impact on the observed bias-$\Delta\chi^2_{(\mathrm{df})}$ correlation. It is clear from this that $P_s(\Delta\chi^2_{(\mathrm{df})}, b)$ also depends to an extent on the choice of $P(\boldsymbol{\theta}_\mathrm{IA})$. We can see that $P(\boldsymbol{\theta}_\mathrm{IA})$ behaves analogously to a prior, restricting the range of possibilities in the subsequent steps. However, given that the purple points in Figure 3.5.1 show a relatively tight correlation and cover a broad range of bias relatively uniformly, we do not expect the details to change things considerably.

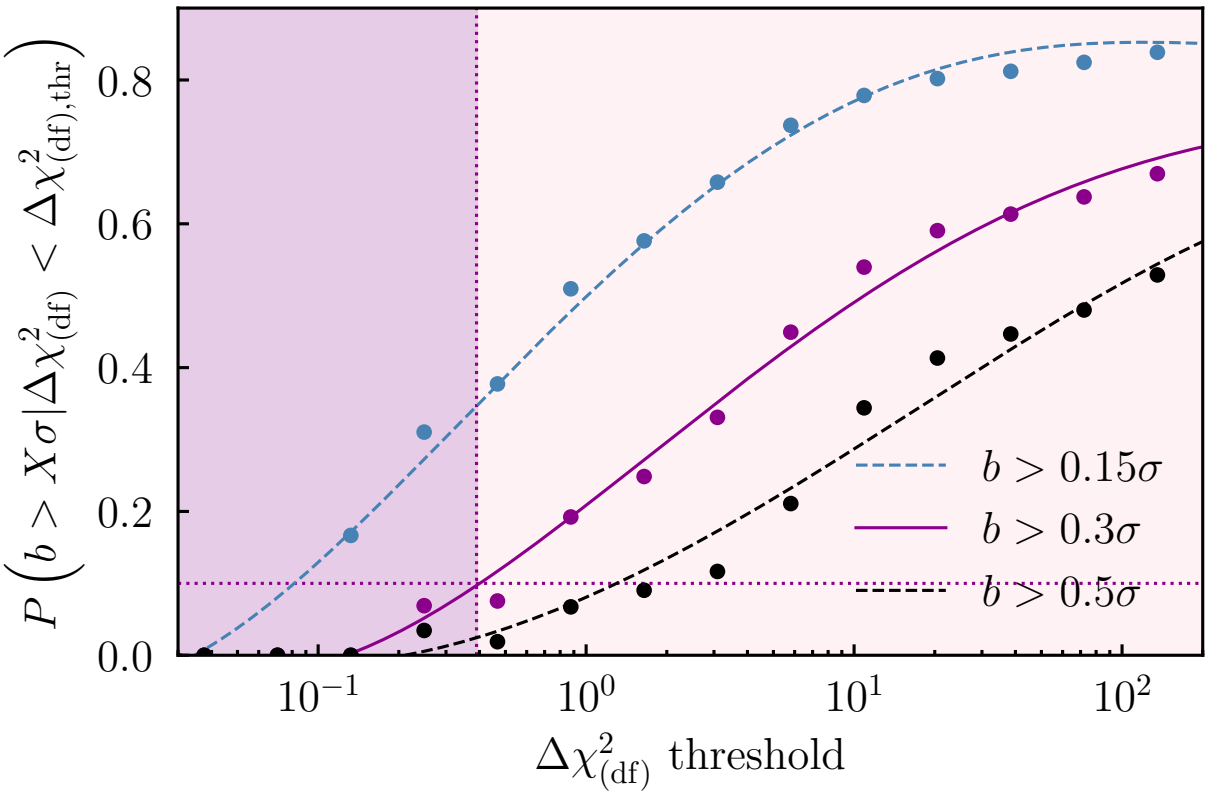

Figure 3.5.3: Probability of exceeding some specified level of cosmological parameter bias, as a function of the threshold $\Delta\chi^2_{\rm (df)}$ value. For a DES Y3-like cosmic shear data vector with unknown noise and IA realisation, and that is found to give an observed $\Delta\chi^2_{\rm (df)}$ lower than threshold the $\Delta\chi^2_{\rm (df)thr}$, P is defined as the probability that the results using NLA are biased by more $X\sigma$ in the $S_8 - \Omega_{\rm m}$ plane. Different values of X are represented by different colours. In each case, we show both the direct measurement of P using importance sampling (coloured points), and the lines are obtained by doing a polynomial fit. For illustrative purposes, we also show the $\Delta\chi^2_{\rm (df)}$ threshold that would guarantee NLA is unbiased to within 0.3σ at a confidence level of 90% (dotted lines and shading).

Table 3.5.1: Confusion Matrix. The samples are split into quadrants, corresponding to the four shaded regions in Figure 3.5.4. The left/right columns show the fraction of IA samples that give a bias above and below 0.3σ. The rows indicate whether or not our method using the calibrated $\Delta\chi^2_{(\mathrm{df})}$ prefers NLA or TATT.

		Model preferred by Bias		
		NLA	TATT	Total
Model preferred by $\Delta\chi^2_{(\mathrm{df})}$	TATT	19.1%	77.1%	96.2%
	NLA	3.4%	0.4%	3.8%
	Total	22.5%	77.5%	100%

The other part of the final sampling of points is the noise distribution $P_N(\widetilde{\Delta\chi^2}_{(\mathrm{df})}, \tilde{b}|\Delta\chi^2_{(\mathrm{df})}, b)$. We obtain this for a particular IA sample by sampling noise realisations, and so transforming $P(\widetilde{\boldsymbol{D}}|\boldsymbol{D}) \rightarrow P_N(\widetilde{\Delta\chi^2}_{(\mathrm{df})}, \tilde{b}|\Delta\chi^2_{(\mathrm{df})}, b)$. This process is again dependent on the covariance matrix, but not on the choice of $P(\boldsymbol{\theta}_{\mathrm{IA}})$ (at least, not directly).

The end result of the above is that, by convolving to get to $P(\widetilde{\Delta\chi^2}_{(\mathrm{df})}, \tilde{b})$, we are able to map out the relationship between a quantity we can measure (the noisy $\widetilde{\Delta\chi^2}_{(\mathrm{df})}$) and the one we are interested in (parameter bias $\tilde{b}$).

Bias Tolerance Implications

We further illustrate our results by taking a concrete example. For our DES Y3 setup, we choose a bias tolerance of $X = 0.3\sigma$, and a bias probability of 10% ($P = 0.1$). Using Figure 3.5.3, this gives us $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}} = 0.4$ (reading across where the horizontal dashed line meets the purple curve), which is shown in Figure 3.5.4 (the horizontal line labelled “empirical threshold"). With the bias and $\Delta\chi^2_{(\mathrm{df})}$ thresholds fixed, the four shaded regions in Figure 3.5.4 distinguish the following possible scenarios: (a) NLA is sufficient (i.e., the bias is below our 0.3σ limit) and $\Delta\chi^2_{(\mathrm{df})}$ chooses NLA (i.e., $\Delta\chi^2_{(\mathrm{df})} < \Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$; purple); (b) NLA is sufficient and $\Delta\chi^2_{(\mathrm{df})}$ chooses TATT (grey); (c) NLA is insufficient and $\Delta\chi^2_{(\mathrm{df})}$ chooses TATT (pink); (d) NLA is insufficient and $\Delta\chi^2_{(\mathrm{df})}$ chooses NLA (red). As we discussed previously, case (d) is the most dangerous, for obvious reasons. Scenario (b) is not ideal (since we may end up with a model that is more complicated than strictly necessary), but does not result in cosmological parameter biases. The different scenarios can be better understood with the help of a confusion matrix, shown in Table 3.5.1.
The columns here represent the model preference according to the amount of bias, and the rows represent the model preference according to the $\Delta\chi^2_{(\mathrm{df})}$. Since we are effectively using $\Delta\chi^2_{(\mathrm{df})}$ as an empirical proxy for bias, we treat the classification according to the latter (i.e., does using NLA cause cosmological parameter biases for a particular data vector exceeding 0.3σ?) as the truth and the label according to the former (i.e., is $\Delta\chi^2_{(\mathrm{df})}$ below $\Delta\chi^2_{(\mathrm{df}),\mathrm{thr}}$?) as the prediction (in machine learning language).

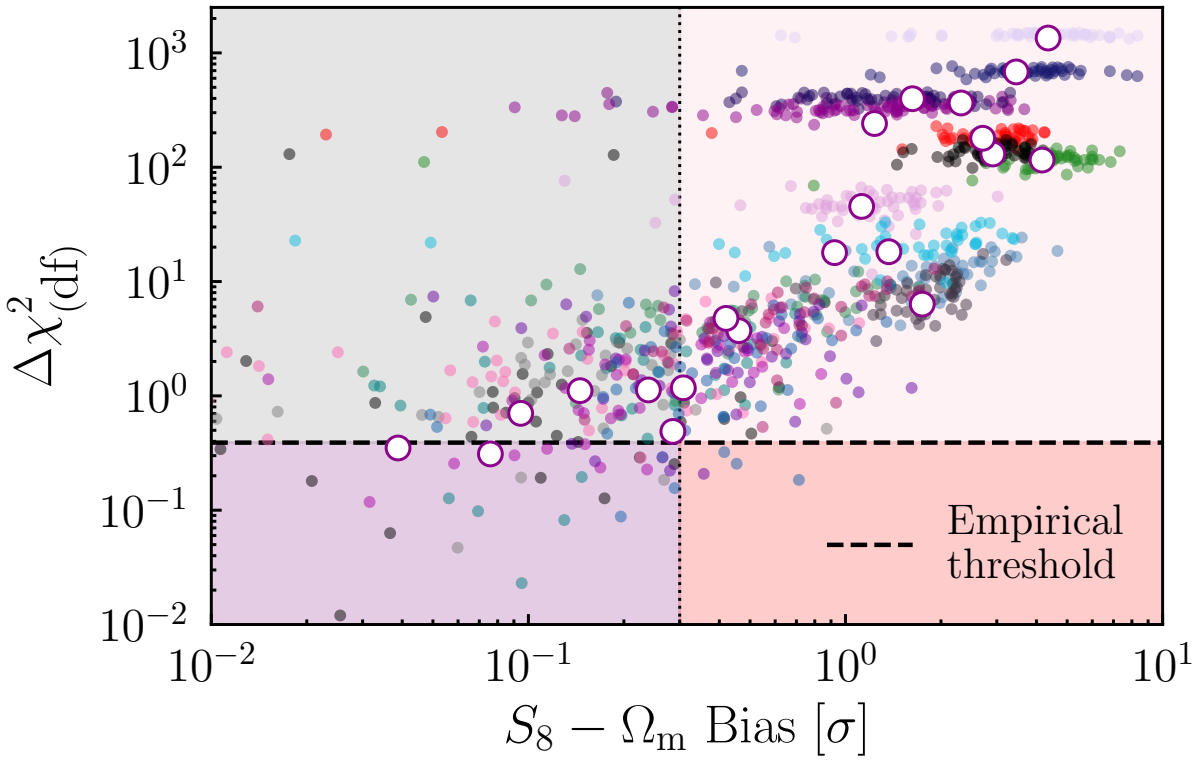

Figure 3.5.4: The impact of data vector noise on $\Delta\chi^2_{(\mathrm{df})}$. The points are the same as in Figure 3.5.2. The horizontal line represents an empirical $\Delta\chi^2_{(\mathrm{df})}$ threshold, derived to ensure bias below 0.3σ with 90% confidence. The four different shaded regions distinguish the following possible scenarios: purple - NLA is sufficient and the calibrated $\Delta\chi^2_{(\mathrm{df})}$ favours NLA ; grey - NLA would be sufficient and yet $\Delta\chi^2_{(\mathrm{df})}$ chooses TATT ; pink - NLA is insufficient and $\Delta\chi^2_{(\mathrm{df})}$ favours TATT ; Red - NLA is insufficient and yet $\Delta\chi^2_{(\mathrm{df})}$ still chooses NLA. This last case is the most dangerous, and the $\Delta\chi^2_{(\mathrm{df})}$ threshold is chosen to keep the fraction of points in this quadrant acceptably small.

We see that in our samples, the cosmological parameter bias indicates that NLA should be preferred about 22.5% of the time, while TATT should be preferred 77.5%. In other words, NLA introduces a bias smaller than our threshold in $\sim 1/4$ of the cases. Note that this fraction is somewhat dependent on our particular choices. A different choice of posteriors in Figure 3.3.1, for example, could change this fraction. We do not, however, expect this to affect the validity of the method.

When it comes to the performance of $\Delta\chi^2_{(\mathrm{df})}$ in identifying the correct model, we see that it favours TATT in 96.2% of the cases, and NLA in only 3.8%. We can see that our method is quite conservative, in the sense that there is a non-negligible false positive rate. That is, it prefers TATT over NLA in 19.1% of cases, even though NLA would not introduce bias to the model above the 0.3σ threshold. Reassuringly, however, we also see that our method is is highly effective in ruling out real bias. The strongest feature of our approach, perhaps, is the fact that it is very unlikely to select NLA if it is, in fact, introducing biases to the analysis. We can see this by the very small population of points in the lower right of the matrix (and the red shaded quadrant in Figure 3.5.4): this happens in only $\sim 0.4\%$ of cases. Put another way, *if* the calibrated $\Delta\chi^2_{(\mathrm{df})}$ favours TATT, there is a roughly 20% chance (19.1/96.2) that NLA would, in reality, have been fine. Conversely, *if* it prefers NLA, there is $\sim 10\%$ (0.4/3.8) that NLA is insufficient. Therefore, even though the end result is somewhat cautious (in that there is a moderate false positive rate for TATT), on the positive side we can be confident that if NLA is in fact preferred by the data, it is very unlikely that it will introduce biases to the analysis. As a remark, however, it is important to acknowledge that a possible conclusion from these results is that simply using the most general model is the cheapest alternative from the perspective of computational resources. It is not obvious that this will always be the case, however, given the dependence on analysis setup and other factors.

It is also worth noticing that although the above discussion applied for our specific choices, we can control the conservatism to a significant degree through our analysis choices. We chose a specific bias tolerance and probability that we considered realistic. By changing these values (for example, allowing a bias probability of 20%, or 25%) one can effectively shift the position of the cross in Figure 3.5.4, and trade off false positives for false negatives. This is another advantage of the method: it makes the level of conservatism explicit (and indeed quantifiable), and allows one to adjust that level as preferred. This is much less true when using alternative approaches to model selection.

3.5.3 A Simpler Approach: How Much Can We Tell From A Single Model?

It is also worth taking a moment to consider a related question: if the true IA scenario is extreme enough to give significant cosmological biases, would there be clear red flags from NLA alone, assuming that no fits were carried out with TATT? If this were the case, it would provide a simpler route – instead of performing empirical calibration using simulations, we could simply run one model on the data, and interpret results to see if a more sophisticated model is needed. Considering our 21 IA scenarios with a fixed noise realisation, we find that around 50% of cases with bias $> 0.3\sigma$ have χ^2 values that appear entirely consistent with being drawn from the corresponding

χ^2 distribution[5], $p(\chi^2_{\rm NLA}) > 0.05$. A similar picture is seen when we consider a single high bias IA scenario with alternative noise realisations – computing $p-$values for each realisation, the majority are above 0.05, even in the presence of bias $> 0.3\sigma$. In other words, even in cases where NLA is significantly biased, it is not necessarily obvious from considering the uncalibrated value of $\chi^2_{\rm NLA}$ alone. In contrast, the method we propose, using $\Delta\chi^2_{\rm (df)}$, can correctly identify the need to use TATT to achieve sufficiently unbiased results $77.1/77.5 \sim 99\%$ of the time (see the confusion matrix in Section 3.5.2).

Likewise, although extreme biases do tend to distort the shape of the posteriors, this is not always true in more moderate (but still significantly biased) cases. Figure 3.5.5 shows the NLA posteriors in a few different IA scenarios, spanning the range from almost no bias (purple shaded), to $\sim 1\sigma$ (pink, open contours). For reference we also show the posteriors from TATT fits to the same data vectors in Appendix 3.B. Taken in isolation, none of these show clear signs of problems with the model. It is also interesting that IA mismodelling bias does not always translate into significantly non-zero values for the inferred NLA parameters. In the medium bias case, for example, A_1 and η_1 are both consistent with zero to $< 1\sigma$. Here there is a relatively strong degeneracy between A_1 and S_8, allowing both $A_1 \sim 0$ combined with low S_8, but also a stronger IA amplitude ($A_1 \sim 1$) with a larger S_8. In projection, this results in broad contours on both parameters (notice the black contours in the upper panel of Figure 3.5.5 are slightly wider than the others, having a longer tail to low S_8 than the pink and purple ones).

3.5.4 Intrinsic alignment modelling & wider implications for weak lensing

The results discussed so far have a number of direct implications for the question of intrinsic alignment model selection. Primarily, we have shown that it is *possible* to perform empirical model selection with lensing data. There is a clear relation between cosmological parameter bias and $\Delta\chi^2_{\rm (df)}$, which allows one to define a threshold that can then be applied to the real data. That said, the failure of conventional statistical metrics (e.g., $p-$values) to identify scenarios with significant cosmological parameter biases is notable, and should be kept in mind when trying to understand statistics derived from any single run on real data. The properly calibrated model statistics, however, provide an alternative to the model selection exercises used in previous analyses, which have tended to rely on either simulated analyses (Secco, Samuroff et al. 2022 Section A3), or arguments based on direct-detection studies (Hikage et al. 2019; Joachimi et al. 2021a, Sections 5.4 and 2.4 respectively). The empirical method is arguably an advance on both; first of all, it avoids questions about what constitutes an "extreme" model, which tend to arise in the simulation-based approach. Since the current best constraints on TATT model parameters are relatively weak, it is relatively easy to select an IA scenario that is both consistent with observations, and which would cause significant bias in an NLA analysis (note that this is still true in light of the most recent DES Y3 results Secco, Samuroff et al. 2022; Amon et al. 2022; DES Collaboration 2022). Our empirical approach also avoids the uncertainties that are inherent in

[5]Where the NLA $p-$value here is calculated by assuming $\chi^2_{\rm NLA}$ is drawn from a χ^2 distribution with 224 degrees of freedom (see Secco, Samuroff et al. 2022). The null hypothesis in this case is that the NLA model is adequate to describe the data, and so small $p-$values would indicate model insufficiency.

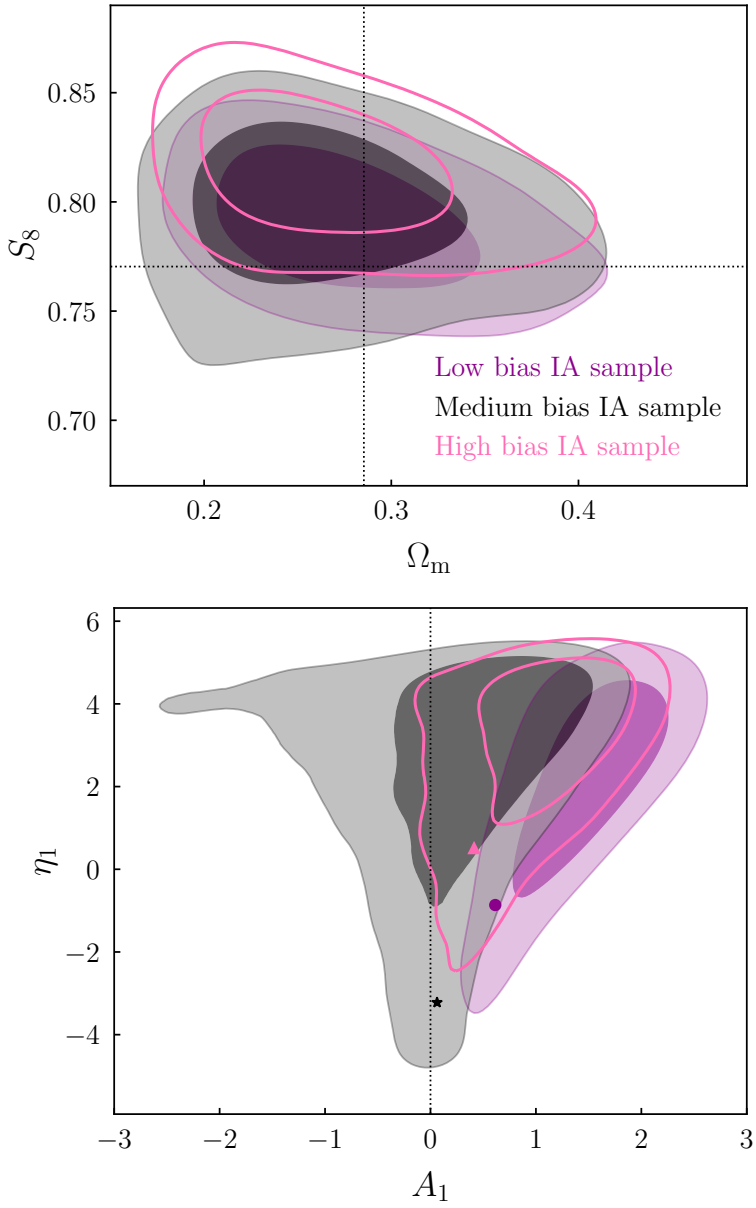

Figure 3.5.5: Examples of the simulated NLA posteriors from three particular IA scenarios with our fiducial noise realisation. These samples were chosen to span a range of bias levels (as defined relative to the TATT posteriors from the same data vectors). In order of severity, the low bias case (purple) has a bias in the $S_8 - \Omega_m$ plane of $\sim 0.1\sigma$, and $\Delta\chi^2_{(df)} = 0.24$, $R = 21.9 \pm 6.1$; the medium bias case (black) has 0.36σ bias, $\Delta\chi^2_{(df)} = 0.49$, $R = 1.5 \pm 0.3$; the high bias case (pink, open) has 0.82σ bias, $\Delta\chi^2_{(df)} = 1.98$, $R = 1.1 \pm 0.2$. The input cosmology and IA parameters are shown as a dashed cross in the upper panel and as coloured points in the lower panel. Notice that the contours are not centred on the input due to the fact that these are noisy data vectors. In all cases the posteriors are not visibly distorted (although in the medium and high bias cases, the η_1 posterior is cut off slightly by the upper prior edge at $\eta_1 = 5$).

extrapolating observations based on direct IA measurements on one very specific type of galaxies to weak lensing measurements on another population entirely.

Although the empirical approach has various strengths, is worth reiterating that data vector noise is a significant source of scatter in the bias-$\Delta\chi^2_{(\mathrm{df})}$ relation. For this reason it is important to accurately simulate the noise properties of the particular data set. While this is in principle simple, given an accurate covariance matrix, it does mean the model selection exercise needs to be repeated for any new data set or changes to the analysis. It also means it is crucial to have a fast and accurate way of estimating posteriors for a large number of noise realisations, such as the IS framework used here.

It is also interesting, finally, to consider how our findings relate to the real Y3 results. Although comparing with the full 3×2pt results is difficult, for the reasons given above, Secco, Samuroff et al. (2022) (Section VIIB and Table III) present a comparison of IA models without shear ratios, an analysis configuration that matches ours. Specifically, comparing the 2 parameter NLA model with 5 parameter TATT, they find $R = 1.70 \pm 0.36$ and $\Delta\chi^2_{(\mathrm{df})} = 5.2/2 = 2.6$. Interpreted with the help of Figure 3.5.3, this puts the risk of NLA being biased by $> 0.3\sigma$ at somewhere around 30%, meaning runs using NLA on Y3 were more likely than not to be unbiased to within the 0.3σ threshold.

3.6 Conclusions

In this paper, we explore the idea that model selection for cosmological analyses could be performed *a posteriori*, being informed by the blinded data themselves. Our goal is to select a model that is sufficient to describe the data, resulting in unbiased parameter constraints at some specified tolerance level. We chose to focus on a specific problem: how best to decide on an intrinsic alignment model for a cosmic shear analysis. This is an important question, and one that has been the subject of much discussion within the weak lensing community in recent years. That said, the basic concept behind our method is much more general, and could be applied in a variety of different contexts; in principle it requires only that the data (including its noise) can be readily simulated.

Using simulated noisy DES Y3 weak lensing data, we tested the method, and identified statistical tools with which to implement it. The main conclusions of our study are as follows:

- We showed a clear relation between the χ^2 difference between two models, and model insufficiency bias on cosmological parameters. This relation was seen to extend across a wide range of biases, from low to high, allowing one to define an empirical $\Delta\chi^2$ threshold in order to ensure bias is below an acceptable level.

- We tested a number of common χ^2-based metrics such as AIC, BIC and pre-defined $p-$value cutoffs. These were seen to be generally under-cautious, favouring the simpler model even in the presence of $1-2\sigma$ parameter biases. This result motivates us to use an empirical $\Delta\chi^2$ calibration. Similarly, when trying to interpret the goodness of fit statistic from a single model, a standard $p(\chi^2) = 0.05$ cutoff is not reliable to rule out significant biases.

- In addition to maximum likelihood-based statistics, we also consider the Bayes factor as a model selection tool. Although useful in the extreme cases, it was seen to be only weakly discriminating for cosmological parameter biases in the range $0.2 - 1\sigma$. We therefore recommend it be used with caution, ideally in conjunction with other model selection metrics.

- Noise is seen to have a potentially significant impact on both the cosmological bias, and the $\Delta\chi^2_{(\mathrm{df})}$, for any given input IA scenario. At high bias, the picture is relatively stable; noise cannot, in general, cause model selection metrics to prefer the simpler model in a case where adopting that model induces large cosmological parameter biases. The reverse is, however, possible. Due to noise, one can end up in scenario with small cosmological parameter bias, but with selection metrics favouring the more complex model. In this regard, our method tends to err on the side of caution.

Although our qualitative findings are general, it is worth bearing in mind that the details are specific to the DES Y3 cosmic shear only setup. Factors such as choice of two-point statistics, covariance and scale cuts could very easily have an impact, as could modelling choices (baryonic treatment, power spectrum, cosmological model etc) and the choice of sampler. It is therefore important that the simulated analyses used to derive a $\Delta\chi^2$ threshold are as close as possible (and ideally identical) to the real setup that will be applied to the blinded data.

Model selection is an important topic in cosmology, and in science more generally. It is quite common to have a set of models under consideration, with little prior knowledge about the values of their parameters; what level of complexity is sufficient to describe the data, given its precision, depends on the unknown true model and its unknown parameter values. Given these circumstances, arguably the most cautious approach would be to use the most flexible model, which is more likely to be unbiased. This paper sets out an alternative method, which allows information in the data to inform model selection. Although applicable in similar situations to Bayesian Model Averaging (BAM; Liddle et al., 2006b; Vardanyan et al., 2011), i.e., where there is not enough prior information to justify choosing one model over another, our approach has the advantage of simplicity, and maintains the idea of a fiducial model, which is often useful for practical purposes. It also avoids the prior dependence of methods such as BAM, which is well documented in the literature. Given its generality, simplicity, and the relatively low level of resources required, we foresee applications of the empirical method discussed in this paper to future analyses as a model selection tool in many contexts.

Data Availability

All simulated data vectors, PolyChord chains and Importance Sampling noise samples used in this work are publicly available at `https://github.com/AndresaCampos/empirical_model_selection`.

Acknowledgements

We thank Scott Dodelson, Sukhdeep Singh, Lucas Secco, Alex Amon, Judit Prat, Agnès Ferté and Jonathan Blazek for useful discussions contributing to this work. Many thanks also to Jessie Muir, Noah Weaverdyck, Otávio Alves, Shivam Pandey, and Cyrille Doux for support with code, in particular with setting up the importance sampling pipeline used in this paper. We finally thank our anonymous reviewer for their helpful comments on the paper. Contour plots were made using the GetDist package (Lewis, 2019).

Andresa Campos thanks the support from the U.S. Department of Energy grant DE-SC0010118 and the NSF AI Institute: Physics of the Future, NSF PHY-2020295. Simon Samuroff is partially supported by NSF grant AST-2206563. RM is supported in part by the Department of Energy grant DE-SC0010118 and in part by a grant from the Simons Foundation (Simons Investigator in Astrophysics, Award ID 620789).

Appendix

3.A Parameters & Priors

Our setup matches the fiducial choices of the DES Y3 cosmic shear analysis. The only significant difference is that, for the sake of simplicity, we choose not to use the additional shear ratio likelihood included by Secco, Samuroff et al. (2022); Amon et al. (2022) (a similar decision was made for validating the analysis choices pre-unblinding; see Krause et al. 2021). As a result, our model space is slightly smaller, since we do not need to vary parameters for galaxy bias or lens photo$-z$ error. The corresponding parameters and their priors are shown in Table 3.A.1. Note that these are *almost* identical to the priors used in the Y3 analysis, except for those on the shear calibration parameters, which have been shifted to match the input to the simulated data.

In a particular setup, one should expect some level of projection effects in the marginal parameter constraints (Krause et al., 2021). Since such offsets are artefacts of the way we choose to visualise our results (i.e., the global best fit is still accurate) it is not, in general, useful to think of them as a form of bias; our method does, however, rely on our ability to interpret differences in the 2D projected $S_8 - \Omega_\mathrm{m}$ plane. It is thus helpful to try to quantify such effects in our case. In Figure 3.A.1 we show the results of our NLA and TATT analyses on an NLA-only data vector, with our fiducial noise realisation. That is, in this case, both the TATT and NLA models can reproduce the data exactly (up to noise). The offset between the best-fitting parameters when using the two models, shown in Figure 3.A.1, is at the level of $\sim 0.1\sigma$. This effectively provides a floor to the bias in our analysis. Although we can occasionally find biases below this level due to noise (see Section 4.5 for discussion), we should consider all these cases as unbiased, at least to within the uncertainty due to projection effects. Note that this is consistent with the results of Krause et al. (2021), who performed a similar test using noiseless data (see their Figure 4).

Note that projection effects are complicated, and may be a function of (among other things) the choice of input parameters, noise realisation, priors and constraining power of the data. Although it is reassuring that our result matches that of Krause et al. (2021), there may still be some residual uncertainty in the size of the effect. This is not, however, necessarily a problem for our method. Indeed, variable projection effects would simply add an extra source of noise in the bias-metric relation, which would be factored into our results in the same way as, e.g., chain-to-chain sampler noise.

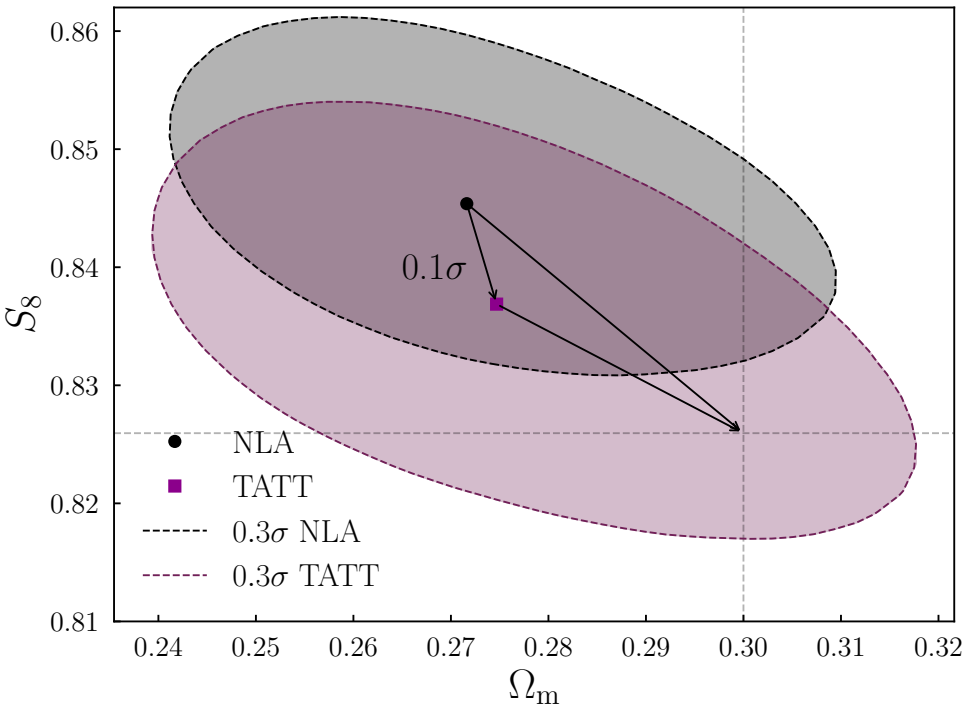

Figure 3.A.1: Projected 0.3σ contours from NLA and TATT chains run on a noisy NLA data vector (see Section 3.4.1 for definitions). The NLA input parameters are $A_1 = 0.7$, $\eta_1 = -1.7$. Since, by construction, both IA models are sufficient to describe the data, any residual offset is thought to be the result of projection effects. As labelled, this is at the level of 0.1σ for our analysis setup.

Table 3.A.1: A summary of the central values and priors used in our analysis. The top seven rows are cosmological parameters, while those in the lower sections are nuisance parameters corresponding to astrophysics and data calibration. Priors are either uniform (U) or normally-distributed, $\mathcal{N}(\mu, \sigma)$. Note the IA parameters are marked with a star because many different values are used as input to our data vectors, as discussed in Section 3.3.2. The values shown here are used for convenience, whenever it is useful to show/discuss a single data realisation (e.g., in Figure 3.3.2).

Parameter	Fiducial Value	Prior
Cosmological Parameters		
$\Omega_{\rm m}$	0.29	$\mathrm{U}[0.1, 0.9]$
$\Omega_{\rm b}$	0.052	$\mathrm{U}[0.03, 0.07]$
h	0.75	$\mathrm{U}[0.55, 0.91]$
$A_{\rm s}$	2.38×10^{-9}	$\mathrm{U}[0.5, 5.0] \times 10^{-9}$
$n_{\rm s}$	0.99	$\mathrm{U}[0.87, 1.07]$
$\Omega_\nu h^2$	0.00053	$\mathrm{U}[0.6, 6.44] \times 10^{-3}$
Calibration Parameters		
m_1	0.0	$\mathcal{N}(0.0, 0.0059)$
m_2	0.0	$\mathcal{N}(0.0, 0.0042)$
m_3	0.0	$\mathcal{N}(0.0, 0.0054)$
m_4	0.0	$\mathcal{N}(0.0, 0.0072)$
Δz_1	0.0	$\mathcal{N}(0.0, 0.018)$
Δz_2	0.0	$\mathcal{N}(0.0, 0.015)$
Δz_3	0.0	$\mathcal{N}(0.0, 0.011)$
Δz_4	0.0	$\mathcal{N}(0.0, 0.017)$
Intrinsic Alignment Parameters *		
A_1	0.7	$\mathrm{U}[-5, 5]$
A_2	-1.36	$\mathrm{U}[-5, 5]$
η_1	-1.7	$\mathrm{U}[-5, 5]$
η_2	-2.5	$\mathrm{U}[-5, 5]$
$b_{\rm TA}$	1.0	$\mathrm{U}[0, 2]$

3.B NLA & TATT Posteriors

For completeness, in Figure 3.B.1 we show the TATT model posteriors for the IA scenarios discussed in Section 4.5 and Figure 3.5.5. In that section we discussed three sets of IA model parameters that were selected to give a range of severity of $S_8 - \Omega_m$ bias in NLA. Our results there showed that significant biases can be present in an NLA analysis without necessarily distorting the shape of the contours or giving a "bad" χ^2 (interpreted in the conventional way, using statistically-motivated cut-offs). As expected, the cosmological parameter contours in Figure 3.B.1 (upper panel) are consistent with each other. Since the data vectors contain (the same) noise, they are offset from the input point slightly. Depending on the input scenario, the width also varies slightly, primarily due to the tail in $A_1 - A_2$ space, which correlates with S_8.

It is worth also briefly commenting here on the shapes of the TATT posteriors. It has been observed before that the TATT model can give rise to teardrop shaped, sometimes slightly bimodal contours in the $A_1 - A_2$ plane (see for example Secco, Samuroff et al. 2022 Fig. 8 and Sánchez, Prat et al. 2022b Fig. 15). A significant tail to positive (or negative) A_2 tends to create a tail in S_8, of the sort seen in grey contours in the bottom panel of Figure 3.B.1. Note however that the shape and asymmetry depends quite heavily on where the posteriors sit in parameter space (meaning the noise realisation as well as the "true" TATT parameters), and on the constraining power of the data (more constraining power tends to trim away some of the non Gaussian tails). It is not clear that we can use any sort of qualitative assessment based on the TATT posterior shape as an indicator for bias in simpler models.

3.C Bayes Ratio

Although we ultimately choose to use the $\Delta\chi^2_{(\mathrm{df})}$ as our model comparison statistic, it is also useful to consider other commonly used alternatives. The Bayes ratio has become a popular tool in weak lensing cosmology in recent years, in part because it in principle contains more information than the likelihood. It is also readily available as the by-product of running a nested sampling algorithm to estimate posteriors.

In Figure 3.C.1 we show the same 21 IA scenarios as in Figures 3.5.1 and 3.5.3, but now using the Bayes factor as our model comparison statistic. In the top panel, the open points again show the noiseless case, with evidence values computed using PolyChord. As we can see, there is a weak correlation, with low bias scenarios tending to give somewhere between "substantial" and "barely worth mentioning" on the Jeffreys scale (the coloured bands). Interestingly, in none of our IA scenarios, not even in the regime that is functionally unbiased, do we see "strong" evidence for NLA.

The scattered points in the top panel of Figure 3.C.1 show the impact of noise, as estimated using importance sampling. Each colour represents an IA sample, with different points representing different noise realisations. Clearly R is significantly more sensitive to noise than $\Delta\chi^2_{(\mathrm{df})}$, as we can see by comparing Figures 3.C.1 and 3.5.3. We observe essentially two scenarios – when the bias in the $S_8 - \Omega_m$ plane is greater than $\sim 1\sigma$, the Bayes Ratio can tells us that NLA is

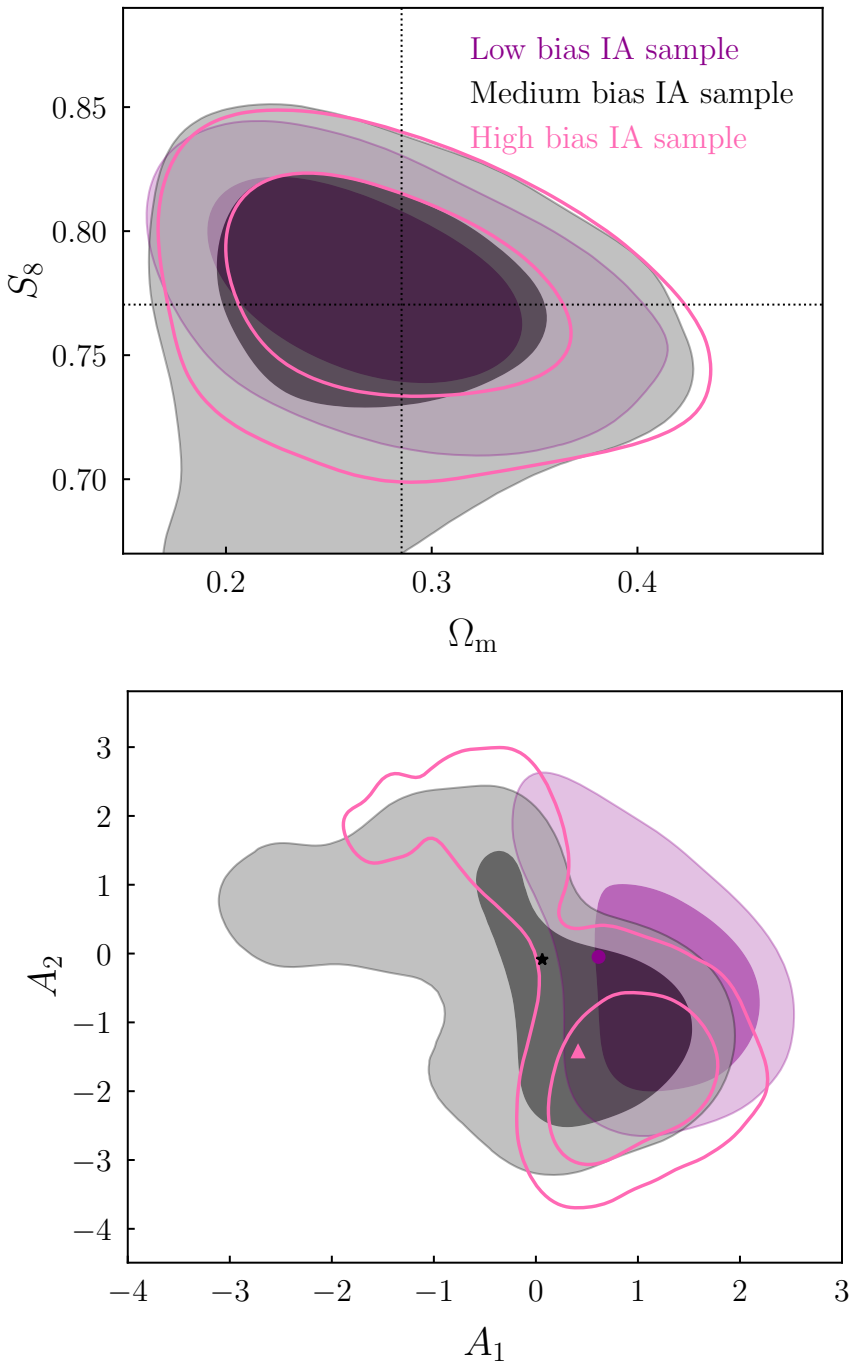

Figure 3.B.1: **Top:** 68% and 95% cosmology confidence contours from TATT model fits on simulated noisy data vectors. Like in Figure 3.5.5, the different colours represent samples selected to cover a range from relatively extreme (i.e., large bias in NLA) to mild (low bias) cases. The dotted cross represents the input cosmological parameters (which is offset from the centre of the contours due to data vector noise). **Bottom:** The same, but showing the two TATT amplitude parameters. The markers (dot, star, triangle) show the input IA parameters for each case.

disfavoured relatively reliably. On the other hand, when the bias in the $S_8 - \Omega_m$ plane is smaller than 1σ, there is a considerable amount of scatter.

The bottom panel in Figure 3.C.1 shows the bias probability, conditioned on the Bayes ratio category. That is, given the data return R in a particular class on the Jeffreys scale, P is the probability that NLA is biased by more than $X\sigma$. There is clearly at least some information here; if the Bayes factor actively favours TATT ($R < 1$), there is a high probability that NLA will be biased, even factoring in noise. Values within the "barely worth mentioning" category could essentially go either way. In the "substantial" category things look better, but even here there is $\sim 10-15\%$ chance of biases more than 0.3σ, and almost 50% chance that NLA is biased by more than 0.15σ in the $S_8 - \Omega_{\rm m}$ plane.

We can perhaps understand the relative noisiness in R by considering Eq. (3.19). Assuming a Gaussian likelihood, the Bayes factor scales as $e^{\Delta\chi^2}$; any small perturbation in χ^2 due to sampling noise will thus be magnified exponentially. We cannot say from this whether this is an inherent issue with the Bayes ratio, or only when estimated using our method of importance sampling. In the absence of an alternative fast method to estimate R for many noise realisations, however, we recommend $\Delta\chi^2_{\rm (df)}$ as a more robust metric to use with our method.

3.D Sampler Comparison

In this appendix we present a brief comparison of two commonly used nested sampling codes: PolyChord and MultiNest. Although a similar (albeit more extensive) exercise is discussed in Lemos, Weaverdyck et al. (2022), their analysis choices differ significantly from ours, and so it is worth revisiting the question. To this end, we re-analyse our 21 noise 0 IA data vectors using MultiNest (500 live points, efficiency= 0.3 , tolerance= 0.01). The results are then compared with our fiducial PolyChord run (500 live points, num_repeats= 30, tolerance= 0.01). We find:

- The two samplers give consistent results for point estimates. That is, both can reliably locate the posterior mean, and the sampling around the peak is comparable, giving a similar level of noise in the best fit. As a result, the χ^2 difference between NLA and TATT analyses is relatively insensitive to the choice of sampler.

- MultiNest is seen to underestimate the width of the 1σ posteriors on cosmological parameters significantly. This is true in both models; combined with the previous point, it leads to a systematic overestimation of the $S_8 - \Omega_{\rm m}$ bias for any given IA scenario. This can be seen in Figure 3.D.1, which shows the posteriors as estimated by both samplers for a particular IA scenario.

- The Bayesian evidence estimates from MultiNest are low compared with PolyChord, as was shown in Lemos, Weaverdyck et al. (2022). Although this is true for both models, the overall result is to increase R (i.e., push the Bayes ratio towards favouring NLA more strongly).

Given the observations listed above, we chose to use PolyChord as our fiducial sampler, despite the runtime advantage of MultiNest.

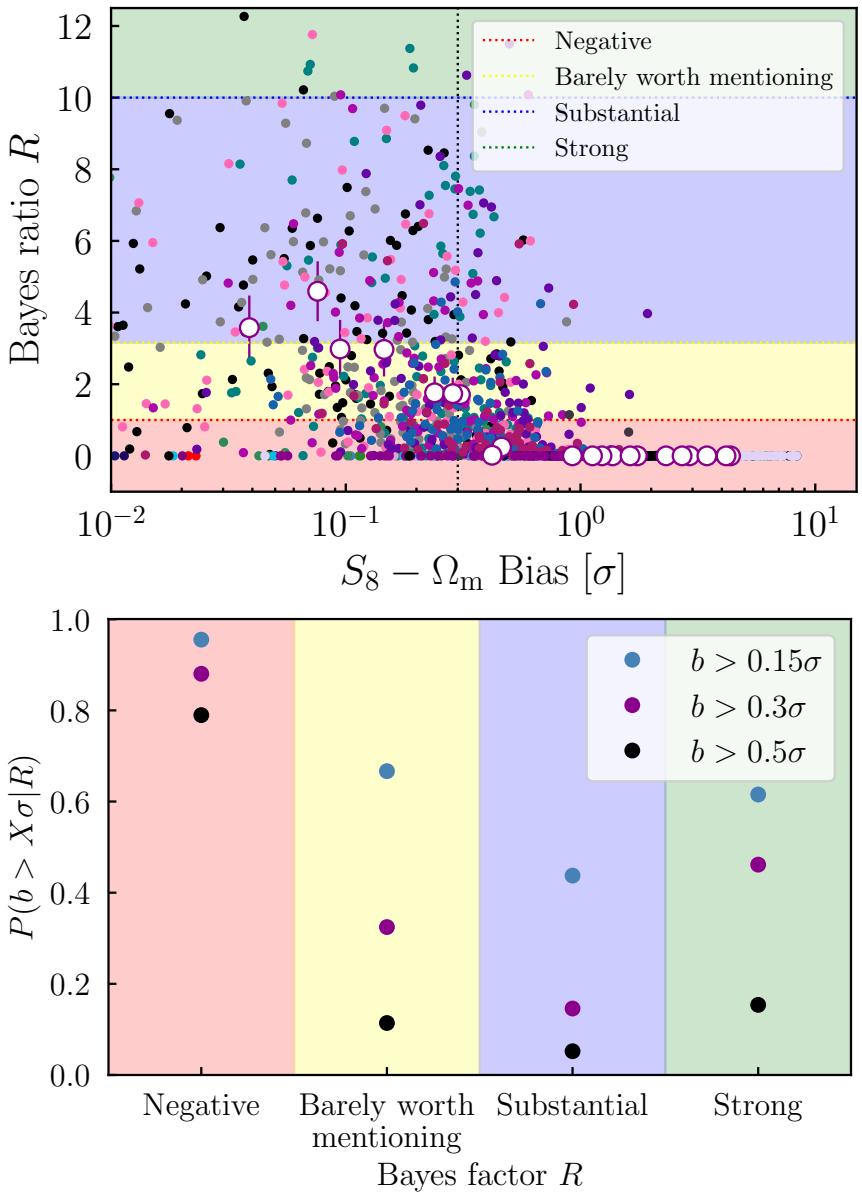

Figure 3.C.1: **Top:** The same 21 samples as in Figure 3.5.1, but now showing the Bayes ratio $R = \mathcal{Z}_{\rm NLA}/\mathcal{Z}_{\rm TATT}$ rather than $\Delta\chi^2_{\rm (df)}$. As before, the open points show the bias/evidence ratios estimated by running PolyChord on noiseless data vectors. The points represent the scatter due to noise (50 noise realisations for each IA scenario; see Section 3.3.3 for details). The vertical line shows the 0.3σ bound, and the shaded colours show how the different ranges of R are interpreted according to the Jeffreys scale. **Bottom:**The same as Figure 3.5.3, but showing the Bayesian factor R (defined as the ratio of Bayesian evidence values obtained from running NLA and TATT on the same data). The coloured bands represent categories on the Jeffreys scale, and P is the probability of more than $X\sigma$ cosmological bias in the NLA model, given an observed Bayes factor in each category.

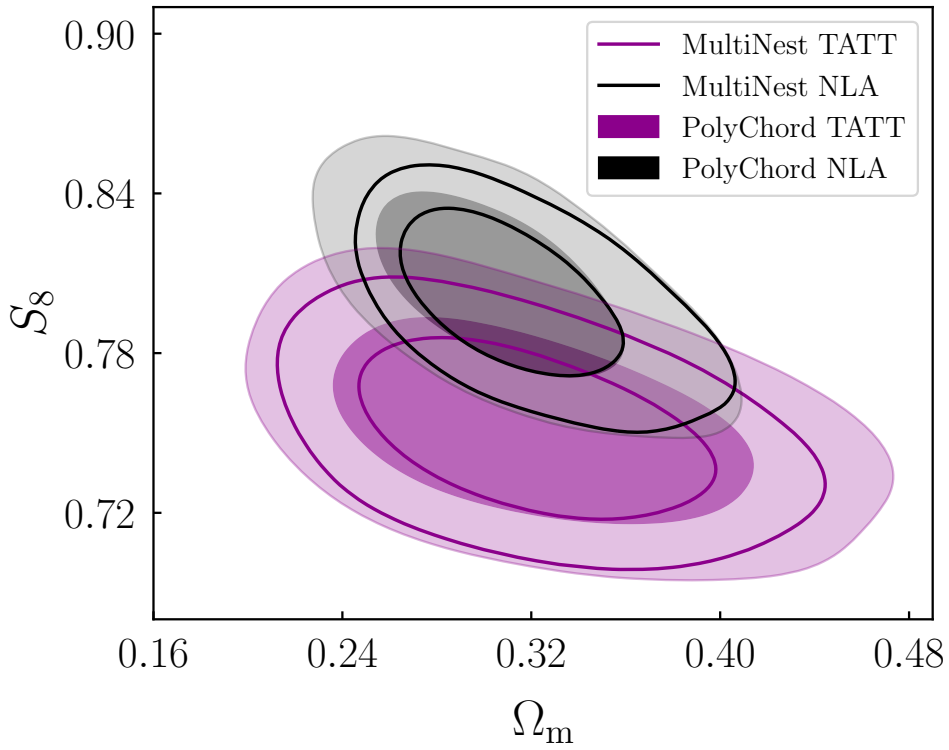

Figure 3.D.1: Marginalised posteriors from a single noisy data vector, with a given input IA scenario. The shaded purple and black contours show the results of fitting that data assuming TATT and NLA respectively, using the PolyChord nested sampling code. The unfilled contours are the results of the same analyses, but using the faster, but less accurate, MultiNest algorithm.

3.E Computational Resources

Here we describe the computational resources used for this paper. The aim is to provide an estimate of the computing power required to apply the method described in Section 3.4.4 to perform model selection. The exact amount of time/resources will naturally vary depending on the details of the analysis pipeline. In our particular case:

- Generate ~ 25 IA samples: less than 1 minute in 1 core.
- Generate ~ 25 IA datavectors: less than 1 minute per datavector in 1 core.
- Run 25 NLA chains using PolyChord: around 22h per chain in 128 cores.
- Run 25 TATT chains using PolyChord: around 28h per chain on 128 cores.
- Generate 50 noise realisations, to be added to each one of the 25 data vectors: less than 1 minute in 1 core.
- Generate importance sampling weights and the χ^2 pool for 25×50 noisy data vectors: around 6-12h on 128 cores.
- Apply the weights to evaluate the NLA and TATT posteriors and compute the bias and best fit in each case: around 6-12h on 128 cores.

Chapter 4

Assessing tension metrics with Dark Energy Survey and Planck data

Abstract

Quantifying tensions — inconsistencies amongst measurements of cosmological parameters by different experiments — has emerged as a crucial part of modern cosmological data analysis. Statistically-significant tensions between two experiments or cosmological probes may indicate new physics extending beyond the standard cosmological model and need to be promptly identified. We apply several tension estimators proposed in the literature to the Dark Energy Survey (DES) large-scale structure measurement and *Planck* cosmic microwave background data. We first evaluate the responsiveness of these metrics to an input tension artificially introduced between the two, using synthetic DES data. We then apply the metrics to the comparison of *Planck* and actual DES Year 1 data. We find that the parameter differences, Eigentension, and Suspiciousness metrics all yield similar results on both simulated and real data, while the Bayes ratio is inconsistent with the rest due to its dependence on the prior volume. Using these metrics, we calculate the tension between DES Year 1 3×2pt and *Planck*, finding the surveys to be in $\sim 2.3\sigma$ tension under the ΛCDM paradigm. This suite of metrics provides a toolset for robustly testing tensions in the DES Year 3 data and beyond.

cosmology: observations – cosmological parameters – methods: statistical

Note on Author Contribution

Despite the authorship order listed in the publication, I wish to clarify my central role in the project described herein. As a co-lead of the Tensions analysis team in DES Y3 together with Pablo Lemos and Marco Raveri, my contributions were instrumental in driving this research forward. The work underpinning this paper relied heavily on simulated data analogous to that from the Dark Energy Survey Year 1, into which controlled artificial tensions in cosmological parameters were introduced – a process I was entirely responsible for. The generation of multiple data vectors and the subsequent Markov Chain Monte Carlo (MCMC) analyses to estimate posterior distributions were under my direct purview. These foundational elements of the project, that required extensive computational work, formed the bedrock upon which the results stand. Upon this groundwork, my

colleagues built further, employing the results I produced to calculate various tension metrics. My role can be accurately regarded as akin to that of a main author, given the breadth and depth of my contributions to the success of the project. This sentiment is shared and endorsed by my fellow co-authors, who acknowledge and agree with the significance of my role in this collaborative endeavor.

4.1 Introduction

Two experiments are generally expected to agree, roughly within the reported errors, on the measured values of cosmological parameters. A disagreement between such measurements — a *tension* — may be a sign of a mistake in one or both analyses, of unaccounted-for systematic errors, or perhaps of new physics. A prominent historical example of such tensions in cosmology is the disagreement between a variety of measurements of the matter density $\Omega_{\rm m}$ in the 1980s and 1990s that was vigorously debated at the time (Peebles, 1984; Efstathiou et al., 1990; Ostriker & Steinhardt, 1995; Krauss & Turner, 1995) and eventually turned out to be explained by the discovery of the accelerating universe (Perlmutter et al., 1999; Riess et al., 1998).

Presently, the discrepancy between the measurements of the Hubble constant using the distance ladder, $H_0 = (74.03 \pm 1.42)$ km/s/Mpc (Riess et al., 2019), and those from *Planck*, $H_0 = (67.4 \pm 0.5)$ km/s/Mpc (Planck Collaboration et al., 2018), is much discussed, as it may be a harbinger of new physics. Similarly, recent measurements of the parameter combination[1] $S_8 \equiv \sigma_8(\Omega_{\rm m}/0.3)^{0.5}$ from large-scale structure by the Dark Energy Survey (DES, DES Collaboration, 2018) and the Kilo Degree Survey (KiDS, Asgari et al., 2020; Heymans et al., 2020) differ from the cosmic microwave background (CMB) estimates from the *Planck* satellite at ~ 2–3σ significance. These $N\sigma$ quantifications of tension are generally understood to correspond to probabilities equivalent to one-dimensional normal distribution, so that 1σ corresponds to 68% confidence that the measurements are discrepant, 2σ corresponds to 95%, etc.

The challenge is how to convert constraints from two data sets into such a probabilistic measure of tension between them. There exist a variety of methods to do this, which are being actively used in the community. While these *tension metrics* are expected to give consistent messages in cases where the two data sets obviously agree or disagree, in more marginal cases the differences amongst them — including how much they depend on an analysis' choice of priors, assumptions of posterior Gaussianity, and the higher-dimensional shape of the posterior —- have the potential to alter the assessment of whether or not two data sets are in agreement.

In the lead-up to cosmological results expected from the analysis of DES year 1 to year 3 data (henceforth simply Y3) and to inform other future cosmological analyses, we wish to provide a comprehensive characterization of how several proposed methods compare to one another. We also wish to confront these results with our intuition for what these metrics ought to be telling us about the agreement or disagreement between measurements. We specifically apply the methods to assess the consistency of DES and *Planck*. This paper complements two earlier analyses that

[1]Here σ_8 is the present-day linear theory root-mean-square amplitude of the matter fluctuations averaged in spheres of radius 8 h^{-1} Mpc.

test the consistency of probes within DES (Miranda et al., 2020; Doux et al., 2021).

These metrics serve only as diagnostics for whether there is tension, and not as a solution. If tension exists, it would indicate either unaccounted-for systematic effects in one or both experiments, or that the underlying model is inadequate to explain the data.

Our basic approach is to create a suite of simulated DES data sets with a controlled level of induced tension relative to the best-fit *Planck* 2018 cosmology. We then apply a number of methods to quantify this synthetic tension and assess their performance. Finally we apply the same tension metrics to quantify any tension between the published constraints from the first year of DES data (DES Y1) and the *Planck* 2015 and 2018 data sets.

The paper is structured as follows: We discuss the difficulties of tension estimation, and present the motivation of the present problem in Sec. 4.2. We then describe our methodology in Sec. 4.3. The different tension metrics studied in this paper are presented in Sec. 4.4. We show results on simulated DES data in Sec. 4.5, apply the tension metrics to DES Y1 in Sec. 4.6, and present our conclusions in Sec. 4.7.

4.2 Motivation

For a tension in a single parameter with an approximately Gaussian posterior distribution, it is easy to define a robust tension metric, as one can just report the one-dimensional difference between the posterior means of the two measurements divided by the quadrature sum of the errors reported by the two experiments. For example, if *Planck* reports that $S_8 = 0.832 \pm 0.013$ (Planck Collaboration et al., 2018) and DES reports $S_8 = 0.782 \pm 0.022$ (Troxel et al., 2018), then one simply adds the errors in quadrature and reports the two results to be different at the level of

$$\frac{\Delta S_8}{\sigma_{S_8}} = \frac{0.832 - 0.782}{\sqrt{0.013^2 + 0.022^2}} = 2.0 \tag{4.1}$$

standard deviations, that is, they are in tension at the 2σ level. However, as soon as we consider a tension in two or more parameters, this simple procedure becomes inadequate because full two-dimensional information cannot be captured by its one-dimensional projections. Fig. 4.2.1 gives an example showing how this intuition breaks down when the parameter space is multi-dimensional. If one were to judge consistency between the two data sets solely through their marginalized 1D constraints, one would conclude that the two data sets are consistent with each other. However, as evident from the comparison of their full 2D parameter constraints, the two data sets are in strong tension. Further complications arise when, for instance, one or more of the posteriors are non-Gaussian, or when the two posteriors originate from different prior assumptions on the parameters of interest.

There is no unique, universally-accepted method to quantify tension under these complicating circumstances. A variety of methods have been proposed, reviewed and tested (Charnock et al., 2017). Given this array of options, it is not obvious what the best choice is for a given analysis. In order to aid in this determination, in this paper we will describe and study several of these methods in order to compare their performance when applied to DES data. In doing so, we distinguish between two kinds of tension:

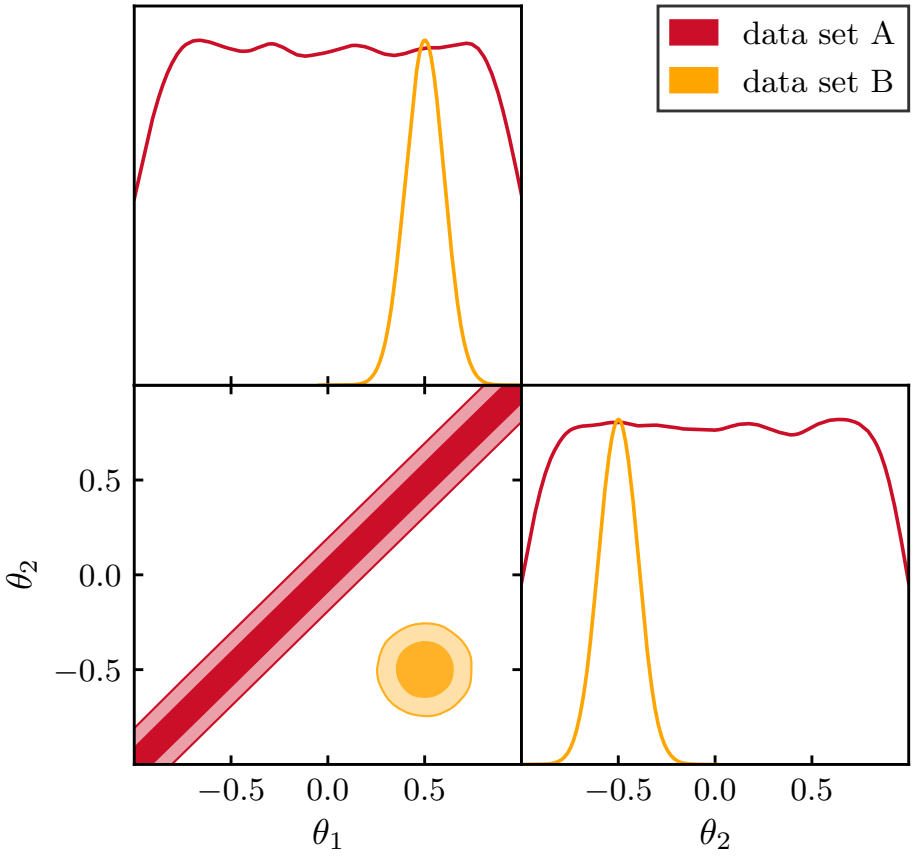

Figure 4.2.1: Toy model example of a set of 2D constraints, where the 1D projections hide the discrepancy between the two data sets. The darker and lighter shade correspond to the 68% and 95% confidence regions respectively.

1. **Internal** tensions, between different cosmological probes within one experiment (e.g. DES cosmic shear vs. galaxy clustering within DES).

2. **External** tensions, between different experiments (e.g. DES vs. *Planck*).

These must be treated differently because data-related systematic effects within the same experiment are often strongly correlated, necessitating use of more complex statistical tools when studying consistency. While our methodology can be applied to either type of tension, here we specifically apply it to the case of external tensions. In addition, we focus on quantifying the tension between the large-scale structure measurements (via the combination of galaxy clustering, galaxy–galaxy lensing and cosmic shear, or often referred to as the "3×2pt" probes) from DES, and the CMB measurements from *Planck*. Internal tension will be separately and additionally studied in Doux et al. (2021) using Posterior Predictive Distributions (PPD) (Gelman et al., 2004), which allow us to quantify tension in the presence of correlated systematic errors in the data, and to visualize the source of tension in the data vector. We do not consider the PPD in this work since it is not well suited to external tensions where there are many parameters that the two data sets do not share.

The challenge of accurately quantifying tension starts to become apparent as we investigate the expected performance of the tension metrics. Naïvely, one might think that shifting one parameter by a controlled number of marginalized N standard deviations would imply that the tension in the full-dimensional space would also be $N\sigma$; or in other words, that the amount of tension in the full, N-dimensional space is equal to the tension projected[2] to the original dimension. However, this is not the case, because of two effects:

- Marginalization can hide tension that can only be seen in higher dimensions. This is caused by the fact that marginalisation leads to loss of information. This means that the full-dimensional tension can be larger than that inferred by looking at 1D distributions of the parameters. This is illustrated with the simple two-dimensional example shown in Fig. 4.2.1: there are two parameters θ_1 and θ_2, and they are highly correlated as measured by experiment 1, but largely uncorrelated as measured by experiment 2. Because experiment 1 determines both parameters separately quite poorly, one-dimensional plots of the posterior show general agreement between measurements of the two experiments. Yet the two-dimensional plot shows that the two contours are significantly separated. This is because the well-measured combination of θ_1 and θ_2 significantly differs between experiment 1 and experiment 2.

- Relatedly, the number of dimensions of the problem also affects the inferred tension. The significance of a difference in parameter estimations between two experiments depends on the number of parameters constrained simultaneously by both experiments. Consider, for example, two experiments that measure the same parameter θ and obtain a one-dimensional 3σ disagreement. The level of significance of this result is much higher if θ is the only parameter constrained by both experiments, than it is if the experiments also measure a

[2] In this paper the terms 'marginalized over' and 'projected' both mean 'integrated over the other parameters'.

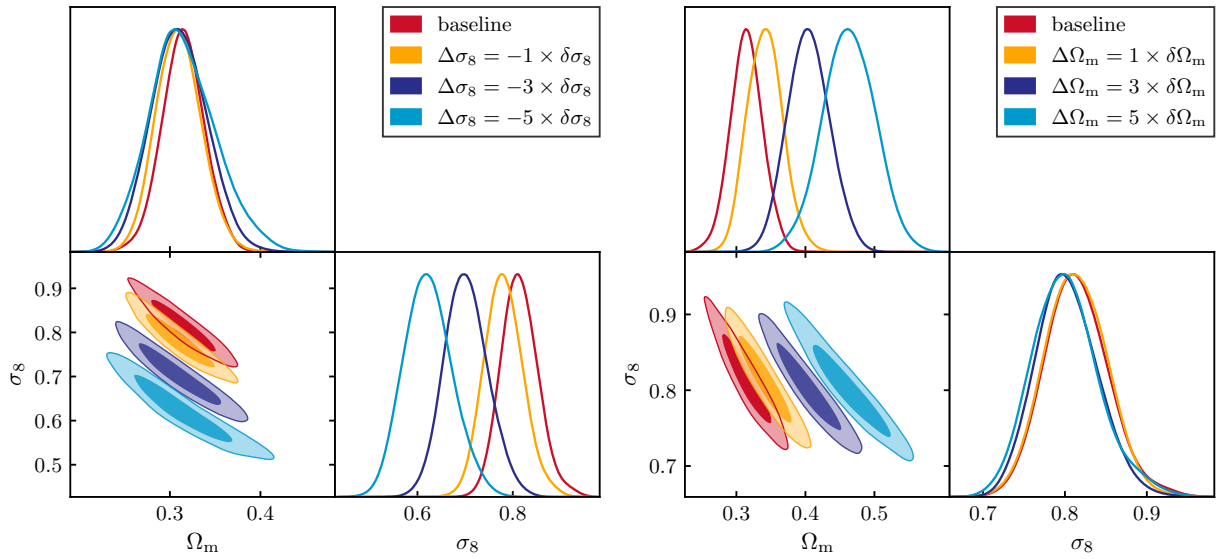

Figure 4.3.1: Marginalized two-dimensional posteriors for some of the simulated DES chains used in this work. The darker and lighter shades correspond to the 68% and 95% confidence regions respectively.

hundred extra parameters, with no significant discrepancies between them. This common problem of the dilution of true tension with multiple comparisons is well known in statistics. For example, Heymans et al. (2020) report a $\sim 3\sigma$ tension with *Planck* in S_8 alone, but a $\sim 2\sigma$ tension when considering the full multi-dimensional parameter space.

4.3 Setting up the problem

The aim of this work is to compare and understand the performance of different metrics for measuring tension between DES and *Planck* constraints on cosmological parameters. If the two experiments report different values for some cosmological parameters, this might be an indicator that their results are not compatible. However, it is important to understand what this discrepancy means when considering the entire model. To do this, we use synthetic DES and *Planck* data sets that have been generated with different input cosmological parameters in order to produce varying levels of expected tension. By applying the various tension metrics to these synthetic data, we can study how they compare to one another and the known input parameter discrepancies. Note that we do not attempt to explain the origin of the possible incompatibility in cosmological parameters reported by two experiments.

We study tension in the context of the flat ΛCDM cosmological model. Our parameters are $\{\Omega_m, \Omega_b, H_0, A_s, n_s\}$, where Ω_m and Ω_b are the density parameters for matter and baryons, respectively; H_0 is the Hubble constant; and A_s and n_s are respectively the amplitude and slope of the primordial curvature power spectrum at a scale of $k = 0.05$ Mpc^{-1}. We assume one massive and two massless neutrino species with the total mass equal to the minimum allowed by

the oscillation experiments, $m_\nu = 0.06$ eV. We do not vary the neutrino mass in our analysis in the simulated data sets, but we do in the reanalysis of tension between DES Y1 and *Planck* of Sec. 4.6, to be consistent with the DES Y1 3×2pt analysis choices (Krause et al.,, 2017). The data and prior choices are further described in Sec. 4.A.

We use the `CosmoSIS` framework[3] (Zuntz et al., 2015) to extract the best-fitting cosmological parameters from the *Planck* 2015 likelihood by sampling it using Nested Sampling (Skilling, 2006), via the `PolyChord` algorithm[4] (Handley et al., 2015a,b). From this chain, we infer the best-fit values of the ΛCDM model parameters according to *Planck* data and use model predictions from these values to generate a baseline simulated DES-like 3×2pt data-vector under the *Planck* cosmology, henceforth referred to as the baseline cosmology. As previously mentioned, the simulated DES data are composed of galaxy clustering, cosmic shear and galaxy–galaxy lensing correlation functions (DES Collaboration, 2018).

4.3.1 Generating a-priori tension

A convenient starting point in our analysis would be synthetically-generated tension in two data sets, corresponding to data vectors generated at different values of cosmological parameters. Precisely how different these two sets of cosmological parameters are should be guided by some preliminary measure of tension. This starting point is henceforth referred to as the "a-priori Gaussian tension", and in this subsection we provide a recipe to define it.

Quantifying the a-priori tension at parameter level with some metrics would make our exercise circular and unfair to other metrics, so it is not a good option. To make progress, we follow a procedure that at least guarantees that the amount of tension we introduce is increasing with increasing shifts, and is, by construction, sensitive to parameters of interest. Using the *Planck* and DES posteriors obtained from their respective baseline data vectors, we first compute the variance in the marginalized one-dimensional posterior distributions for Ω_m and σ_8, referred to as $\mathrm{var}(\theta)$, where $\theta \in \{\Omega_\mathrm{m}, \sigma_8\}$. We then shift each parameter by a multiple of the quantity

$$\delta\theta = \sqrt{\mathrm{var}(\theta_\mathrm{DES}) + \mathrm{var}(\theta_{Planck})} \tag{4.2}$$

and generate simulated DES data vectors with either Ω_m or σ_8 shifted by integer multiples of the corresponding $\delta\theta$. We indicate the total shift with $\Delta\theta \equiv \alpha\delta\theta$ for a given integer α. We then use those data vectors to obtain simulated DES chains. We shift σ_8 towards lower values than *Planck*'s, and Ω_m towards higher values, for simplicity, but we would expect to obtain similar results if the shifts were done in the opposite directions.

A shift in σ_8 is obtained by changing the input value of A_s. Shifting Ω_m, on the other hand, changes the history of structure growth and thereby σ_8; we compensate for this collateral shift in σ_8 by counter-shifting A_s. The DES constraints (shown in the Ω_m–σ_8 plane) from a representative subset of these shifted synthetic data are shown in Fig. 4.3.1.

If we approximate the difference between the *Planck* and DES posteriors as a Gaussian distribution in multiple dimensions we can now ask, *a priori*, what the significance of these shifts

[3] `https://bitbucket.org/joezuntz/cosmosis/wiki/Home`

[4] `https://github.com/PolyChord/PolyChordLite`

is (in the Ω_m–A_s plane) by computing

$$\chi^2 = \delta\theta^T (C_D + C_P)^{-1} \delta\theta \tag{4.3}$$

where C_D and C_P are the 2×2 covariance matrices in (Ω_m, A_s) for DES and *Planck* respectively. Because we are changing only two parameters, the quantity has two degrees of freedom. Note that this is just the generalization of Eq. (4.1) to multiple dimensions. While the Gaussian approximation is not expected to be accurate, especially in the tails of the posteriors, it is expected to be a reasonable guess of the tension that we are inputting into our synthetic examples.

Fig. 4.3.2 shows the distribution of shifted parameter combinations we describe above, as well as the baseline *Planck* + DES parameter constraints. Specifically, the contour shows the combined baseline *Planck* + DES constraints, while the markers show the best-fit values of individual shifted DES-only constraints. We can immediately see that, in multiple dimensions, the tension that we attributed to a one-dimensional shift is higher since Ω_m and σ_8 are correlated.

To quantify the significance of the shifts shown in Fig. 4.3.2, we calculate from Eq. (4.3) the probability to exceed (PTE) our input shifts in the Gaussian case. For example, we would like to associate a 'one-sigma tension' to an Ω_m shift that lies precisely on the edge of the 68% confidence region. We thus adopt a simple 1D Gaussian conversion

$$N_\sigma \equiv \sqrt{2}\,\mathrm{Erf}^{-1}(\mathrm{PTE}), \tag{4.4}$$

where Erf^{-1} is the inverse error function. Given a probability to exceed, N_σ matches that probability with the number of standard deviations that an equivalent event from a 1D Gaussian distribution would have. Note that the conversion in Eq. (4.4) is only a convenient proxy to report high statistical significance results, and does not assume Gaussianity *per se* in any of the statistics.

The resulting evaluation of the a-priori Gaussian tension is shown in Tab. 4.3.1. Here the first column shows the parameter shift applied to DES data in the (Ω_m, σ_8) space, where each parameter is shifted by a half-integer multiple of its reported (marginalized) error. The second column shows the full-parameter-space tension calculated using Eq. (4.4) as described above. Note that the 'input shifts' in Ω_m lead to higher tension than those in σ_8. This is because shifting Ω_m while keeping σ_8 fixed also leads to a shift in A_s, which increases the tension in the full-dimensional space.

Finally, let us note that the a-priori tension, by its construction, does not contain stochastic noise, as it effectively measures the distance in the space of input cosmological parameters. This is in contrast with all of the tension metrics that we study below, which are applied to random realizations of data that do contain noise. The fact that the effectively noiseless input tension is being compared to tension measurements applied on noisy data is one reason why we do not expect a perfect match between the two. We will return to this point in Sec. 4.5.

4.4 Tension Metrics

This section describes the tension metrics that we will be comparing in this work. Several metrics have been proposed for quantifying tension between cosmological data sets. In this work, we

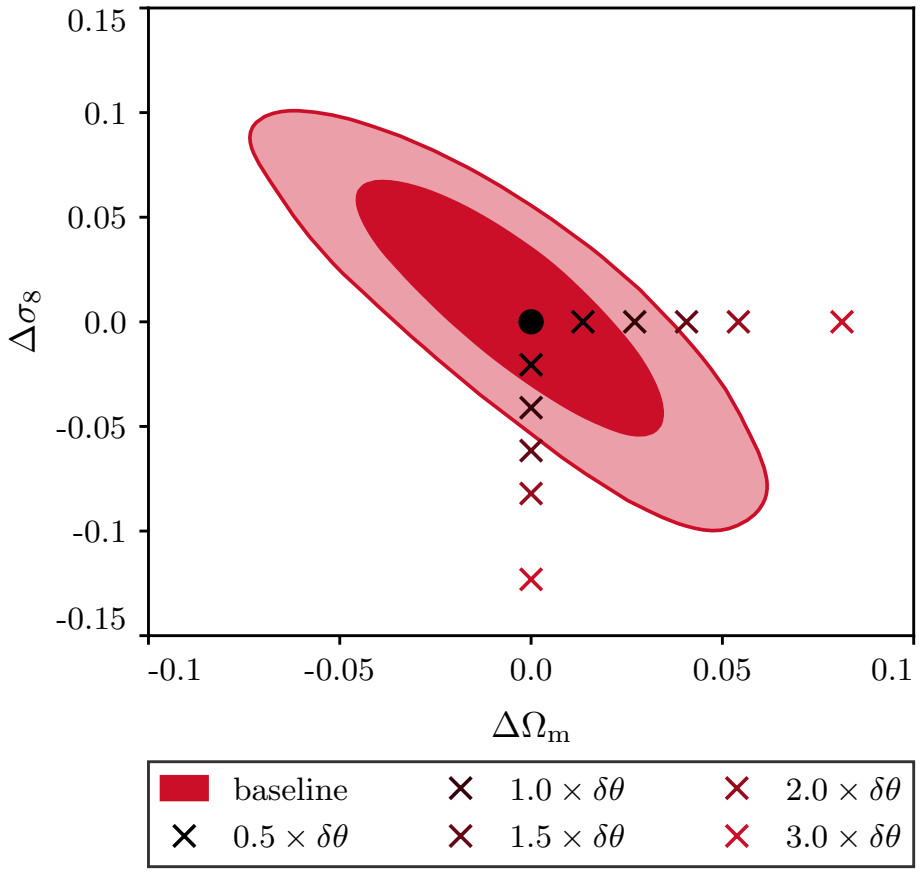

Figure 4.3.2: 68% and 95% confidence regions of the constraint on the differences in parameters as measured by DES and *Planck*, constructed as discussed in Sec. 4.3. The markers indicate the location of the synthetic input shifts. The corresponding a-priori Gaussian tension is shown in Tab. 4.3.1.

Evaluation of a-priori Gaussian tension	
$(\Omega_{\rm m}, \sigma_8)$ shift	full-par-space N-σ
$\Delta\sigma_8 = -0.5 \times \delta\sigma_8$	$0.02\,\sigma$
$\Delta\Omega_{\rm m} = +0.5 \times \delta\Omega_{\rm m}$	$0.09\,\sigma$
$\Delta\sigma_8 = -1 \times \delta\sigma_8$	$0.4\,\sigma$
$\Delta\Omega_{\rm m} = +1 \times \delta\Omega_{\rm m}$	$1.0\,\sigma$
$\Delta\sigma_8 = -1.5 \times \delta\sigma_8$	$1.1\,\sigma$
$\Delta\Omega_{\rm m} = +1.5 \times \delta\Omega_{\rm m}$	$2.3\,\sigma$
$\Delta\sigma_8 = -2 \times \delta\sigma_8$	$2.0\,\sigma$
$\Delta\Omega_{\rm m} = +2 \times \delta\Omega_{\rm m}$	$3.8\,\sigma$
$\Delta\sigma_8 = -3 \times \delta\sigma_8$	$3.7\,\sigma$
$\Delta\Omega_{\rm m} = +3 \times \delta\Omega_{\rm m}$	$> 5\,\sigma$
$\Delta\sigma_8 = -5 \times \delta\sigma_8$	$> 5\,\sigma$
$\Delta\Omega_{\rm m} = +5 \times \delta\Omega_{\rm m}$	$> 5\,\sigma$

Table 4.3.1: Evaluation of a-priori Gaussian tension for controlled shifts in (σ_8 and $\Omega_{\rm m}$). The $\delta\theta$ by whose half-integer value we are shifting these parameters is referring to their respective 1D marginalized posterior as in Eq. (4.2). See Eq. (4.4) for the explanation how we convert these shifts into the "number of sigmas" in the full parameter space, shown in the second column.

select a series of methods that we believe to be appropriate to our data, and which are distinct enough to highlight the strengths and failure modes of each metric. We separate the tension metrics into two subcategories, since while all methods aim to quantify tension between data sets, they answer slightly different questions:

- **Evidence-based methods** seek to answer the question:
  Given hypothesis H_1: 'The assumed model is capable of generating the data observed by both experiments', and hypothesis H_2: 'The assumed model is not capable of generating the data observed by both experiments', which hypothesis is preferred by the data under the assumed model'?

- **Parameter-space methods** seek to answer the question:
  What is the statistical significance of the differences between the posteriors for experiments A and B, within the parameter space analyzed by both experiments?

All of the tension metrics that we consider solve the problems that we have discussed in Sec. 4.2 by considering all dimensions of parameter space. In addition, since they provide results in terms of probabilities, they are independent of the specific parametrizations that are used.

The remainder of this section describes these tension metrics. The results for these metrics will be shown in Sec. 4.5.

4.4.1 Bayesian evidence ratio

The Bayesian evidence ratio, or Bayes ratio R, is an evidence-based method, defined for independent data sets A and B as (Marshall et al., 2006):

$$R \equiv \frac{\mathcal{Z}_{AB}}{\mathcal{Z}_A \mathcal{Z}_B}. \tag{4.5}$$

Here, $\mathcal{Z}_D$ is the Bayesian Evidence, defined as the probability of measuring the observed data D for a given model M, which can be obtained marginalising over all the model parameters θ:

$$\mathcal{Z}_D \equiv P(D|M) = \int \mathrm{d}\theta \; P(D|\theta, M) P(\theta|M). \tag{4.6}$$

Henceforth, we adopt the following notation for Bayes' theorem:

$$\mathcal{P} = \frac{\mathcal{L} \times \Pi}{\mathcal{Z}} \tag{4.7}$$

where $\mathcal{P} \equiv P(\theta|D, M)$ is called the posterior, $\mathcal{L} \equiv P(D|\theta, M)$ is the likelihood, and $\Pi \equiv P(\theta|M)$ is the prior. The Bayesian Evidence is a difficult quantity to calculate, as it requires integrating a probability distribution over a large number of dimensions. One of the most frequently-used tools to calculate Bayesian Evidences is Nested Sampling (Skilling, 2006), which also produces posterior distributions. There exist publicly-available codes for Nested Sampling calculations, such as `MULTINEST` (Feroz et al., 2009) and `POLYCHORD` (Handley et al., 2015a,b).

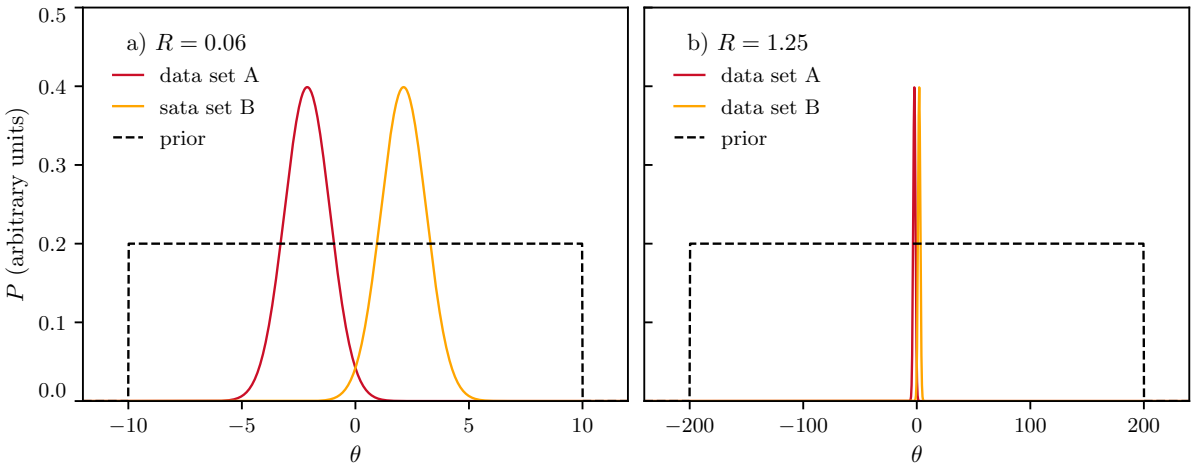

Figure 4.4.1: Example of the prior-volume dependence of R. In amber and red are two gaussians that are at a 3σ tension. The black dotted line is the prior (note that it is not normalized, to make it easier to visualize). When we use a uniform prior in the range $[-10, 10]$ (left panel), R is much smaller than one, which means the data sets are in tension. When we increase the prior to $[-200, 200]$ (right panel), R becomes greater than one, indicating agreement. This example, although extreme, illustrates a possible issue of the Bayes ratio as a tension metric.

$\log R$	Interpretation
> 2.3	Strong agreement
$(1.2, 2.3)$	Substantial agreement
$(-1.2, 1.2)$	Inconclusive
$(-2.3, -1.2)$	Substantial tension
< -2.3	Strong tension

Table 4.4.1: Jeffreys' scale used by (DES Collaboration, 2018) to quantify agreement or tension between data sets (Jeffreys, 1998).

In the Bayes ratio R as written in Eq. (4.5), the numerator requires both data sets to be simultaneously explained by the same parameter values within the model, while the denominator allows each data set to be explained by different parameter values (still within the same assumed underlying model). A more intuitive interpretation (Amendola et al., 2013; Raveri & Hu, 2019; Handley & Lemos, 2019b) uses Bayes theorem to rewrite this as

$$R = \frac{P(A|B, M)}{P(A|M)}\,, \tag{4.8}$$

(where data sets A and B can be interchanged). That is, does the existence of data set B make the data set A more or less likely than it would be in the absence of B, all within the context of assumed model M? Therefore, a ratio of probabilities $R \gg 1$ is interpreted as the data sets being consistent, while $R \ll 1$ indicates that the data sets are in tension. This tension metric has several desirable properties: it is a global statistic (that is, operates on the full parameter space), and it is symmetric between data sets (so tension between data A and data B is the same as tension between B and A). For these reasons, R was used in DES Collaboration (2018), to quantify tension between the DES Y1 measurements and external data sets.

This new interpretation carries an important issue, which is R's dependence on the prior volume: as described by Handley & Lemos (2019b), Eq. (4.5) can be rewritten as:

$$R \equiv \int \mathrm{d}\theta \, \frac{\mathcal{P}_A \mathcal{P}_B}{\Pi}. \tag{4.9}$$

For a flat and uninformative prior, R is therefore proportional to the prior volume. For example, doubling the prior volume doubles the value of R, and increases the agreement between the data sets independently of the shape of the posteriors. As an extreme case, one could increase the prior range arbitrarily to make any two posteriors consistent according to R. This is illustrated by Fig. 4.4.1, which gives two equal-width Gaussians horizontally offset by 3σ. The Bayes ratio is close to zero when the prior encompasses relatively tightly the bulk of the two distributions, but goes up to $R > 1$ if the prior is made sufficiently wide. In the latter case, the Bayes-ratio-logic says that the two Gaussians are close to each other *relative to the width of the prior*, and hence are reported to not be in any tension. This prior dependence is therefore a central feature of the

Bayes ratio. Nevertheless, such a prominent role for the prior may be worrying in situations when physically-motivated priors are not available.

A second concern about the Bayes ratio R is that its raw numerical value needs calibration. R is the ratio of probabilities (see Eq. (4.5)) and one often uses the Jeffreys' scale (Jeffreys (1998); see Tab. 4.4.1) to convert the different outcomes to interpretations about the presence of tension between data sets. However, the boundaries in Jeffreys' scale are arbitrary, and they lack obvious interpretation as a statistical significance.

Both the interpretation and the calibration problem can be circumvented if another tension metric is used to calibrate the Bayes ratio. In this paper, we use the simulated data vectors described in Sec. 4.3 to calibrate the Bayes ratio outcomes (along with those from other tension metrics). Note, however, that this calibration is very specific to our choice of the problem, such as the observables, the parameter space, or the priors we employ. Our results would not be generalizable to an arbitrary cosmological analysis.

4.4.2 Bayesian Suspiciousness

Bayesian Suspiciousness (Handley & Lemos, 2019a) is an evidence-based method, introduced as an alternative to the Bayes ratio from Sec. 4.4.1 for the case of priors which, instead of being motivated by prior knowledge, are purposefully wide and uninformative. This is the case for DES, where wide priors are chosen with the goal of obtaining DES-only constraints. The idea is the following: We divide the Bayes ratio R in two parts, one that quantifies the probability of the data sets matching given the prior width, and another one that quantifies their actual mismatch. The first part is quantified by the information ratio I, defined as:

$$\log I \equiv \mathcal{D}_A + \mathcal{D}_B - \mathcal{D}_{AB}, \tag{4.10}$$

where $\mathcal{D}$ is the Kullback–Leibler Divergence (Kullback & Leibler, 1951):

$$\mathcal{D} \equiv \int \mathcal{P} \log\left(\frac{\mathcal{P}}{\Pi}\right) \mathrm{d}\theta\,. \tag{4.11}$$

The Kullback–Leibler Divergence is particularly well suited to eliminate the prior dependence from the Bayes ratio, as it quantifies how much information has been gained going from the prior Π to the posterior $\mathcal{P}$. Therefore, it encloses the prior dependence that we want to eliminate. The Kullback–Leibler Divergence has been extensively used in cosmology (e.g. Hosoya et al., 2004; Verde et al., 2013; Seehars et al., 2014, 2016; Grandis et al., 2016; Nicola et al., 2019).

The part of the Bayes ratio R that is left after subtracting the dependence on prior volume depends only on the actual mismatch between the posteriors, and it is what we call Bayesian Suspiciousness S:

$$\log S = \log R - \log I. \tag{4.12}$$

As explained in Sec. 4.4.1 and in Handley & Lemos (2019b), the main concern regarding the Bayes ratio R is that the tension can be 'hidden' by widening the priors. S can be understood as the version of R that corresponds to the smallest priors that do not significantly alter the posterior.

It also has two useful qualities that R lacks: It does not depend on the prior volume and, in the case of Gaussian posteriors, it follows a χ^2_d distribution, where d is the effective number of degrees of freedom constrained by both data sets. Therefore, we can assign a *tension probability* p_T as the p-value of the distribution. This tension probability quantifies the probability of the observed tension occurring by chance. While the chi-squared interpretation relies on the approximation of Gaussian posteriors,[5] the rest of this section does not, so the value and sign of S can be used to measure tension for any posterior distributions.

To obtain the value of p_T, we need to calculate the effective number of dimensions constrained by the combination of the data sets. While there are several available methods to do this, we propose using the Bayesian Model Dimensionality (Handley & Lemos, 2019a):

$$d = 2 \int \mathcal{P} \left(\log \frac{\mathcal{P}}{\Pi} - \mathcal{D} \right)^2 . \tag{4.13}$$

This formula is analogous to the more traditional Bayesian Model Complexity (BMC) (Spiegelhalter et al., 2002) used in previous cosmological analyses (e.g. Kunz et al., 2006; Bridges et al., 2009), with which it shares the property that it is formed of Bayesian quantities and recovers a value of $d = 1$ for the 1D Gaussian case. But while the BMC requires the use of either the mean or maximum-posterior parameter values and is hence subject to sampling error (i.e. numerical noise due to a finite length of an MCMC chain), Eq. (4.13) does not suffer from these issues (Handley & Lemos, 2019a).

While the Suspiciousness is according to our definition an evidence-based method, it has been recently shown (Heymans et al., 2020) that it can be reformulated as the difference of the log-likelihood expectation values of joint and individual data sets, leading to a relation between the suspiciousness and the goodness-of-fit loss introduced in Sec. 4.4.5 (Joudaki et al., 2022) through the Deviance Information Criterion (Spiegelhalter et al., 2001) This shows that despite them being defined very differently, there are fundamental relations between these statistics.

All the quantities discussed in this subsection can be simply obtained from a single nested sampling chain (in the case of the BMD, or even an MCMC chain), which means that their computational cost is the same as that of the Bayes ratio introduced in Sec. 4.4.1. Nested sampling can also give us an estimate of the sampling error, by re-sampling the sample weights (Higson et al., 2018). Joachimi et al. (2021a) noted that this method can lead to noise in the dimensionality calculation. This noise was included in this work, and contributes to the error in the estimate of the tension probability. All calculations are implemented in the python package `anesthetic`[6] (Handley, 2019); an example on how to calculate these quantities can be found at `https://github.com/Pablo-Lemos/Suspiciousness-CosmoSIS`.

[5] As pointed out by Handley & Lemos (2019b), non-Gaussian posteriors can be 'Gaussianized' using Box–Cox transformations (Box & Cox, 1964; Joachimi & Taylor, 2011; Schuhmann et al., 2016), that preserve the value of S. Therefore, the chi-squared interpretation of S derived in the Gaussian case can be approximately valid even for posteriors that do not look Gaussian, even if it is not guaranteed that both posteriors can be Gaussianized simultaneously.

[6] `https://github.com/williamjameshandley/anesthetic`

4.4.3 Parameter differences

Another estimator that we consider is the Monte Carlo estimate of the probability of a parameter difference as described in Raveri et al. (2020). This is a parameter-space method, which relies on the computation of the parameter difference probability density $\mathcal{P}(\Delta\theta)$. In the case of two uncorrelated data sets this is given by the convolution integral:

$$\mathcal{P}(\Delta\theta) = \int_{V_p} \mathcal{P}_A(\theta)\mathcal{P}_B(\theta - \Delta\theta)d\theta \tag{4.14}$$

where P_A and P_B are the two parameter posterior distributions and V_p is the support of the prior, i.e. the region of parameter space where the prior is non-vanishing. Notice that this probability density has been marginalized over the value of the parameters and only constrains their difference.

Once the density of parameter shifts is obtained one can quantify the probability that a genuine shift exists:

$$\Delta = \int_{\mathcal{P}(\Delta\theta)>\mathcal{P}(0)} \mathcal{P}(\Delta\theta)\, d\Delta\theta \tag{4.15}$$

which is the posterior mass above the iso-probability contour for no shift, $\Delta\theta = 0$. Note that since Eq. (4.15) is the integral of a probability density it is invariant under reparametrizations.

Equations (4.14) and (4.15) look straightforward, but their evaluation is greatly complicated in parameter spaces with a large number of dimensions. In such cases (which are typical in cosmological applications), the posterior samples cannot be easily smoothed or interpolated to a continuous function, and we are left to work exclusively with N_A samples from the posterior P_A and N_B from P_B, i.e. discrete representations of the posteriors of interest. Each one of the $N_A N_B$ pairs of samples corresponds to one term on the right-hand side of Eq. (4.14) (with $\Delta\theta = \theta_A - \theta_B$, where θ_A and θ_B are the parameter values for that pair).[7]

To make progress, we perform the integral in Eq. (4.15) with a Monte Carlo algorithm. One computes the Kernel Density Estimate (KDE) probability of $\Delta\theta = 0$ and then the KDE probability of each of the samples of the parameter difference posterior. The number of samples with KDE probability above zero divided by the total number of samples is the Monte Carlo estimate of the integral in Eq. (4.15) and the error can be estimated from the binomial distribution. This approach largely mitigates the need for an accurate estimate of the optimal KDE smoothing scale. In practice we use a multivariate Gaussian kernel with smoothing scale fixed by the Silverman's rule (Chacón & Duong, 2018).

We use the implementation of this tension estimator in the `tensiometer`[8] code.

4.4.4 Parameter differences in update form

Another parameter-space method that we consider is the update difference-in-mean (UDM) statistic, as defined in Raveri & Hu (2019). This compares the mean parameters determined from one

[7] In the case of weighted samples the weight of the parameter difference sample is the product of the two weights.
[8] `https://github.com/mraveri/tensiometer`

data set, $\hat{\theta}^A$, with their updated value, $\hat{\theta}^{A+B}$, obtained after adding another data set. The shifts in parameters are then weighted by their inverse covariance to give

$$Q_{\rm UDM} = (\hat{\theta}^{A+B} - \hat{\theta}^A)^T \left(C^A - C^{A+B}\right)^{-1} (\hat{\theta}^{A+B} - \hat{\theta}^A) \tag{4.16}$$

where C^A and C^{A+B} are the posterior covariances of the single data set A and the joint data set $A+B$. If the parameters $\hat{\theta}^A$ and $\hat{\theta}^{A+B}$ are Gaussian distributed then $Q_{\rm UDM}$ is chi-squared distributed with $\mathrm{rank}(C^A - C^{A+B})$ degrees of freedom. These degrees of freedom are the parameters that are measured by both data sets A and B and are the only ones that can actively contribute to a tension between the two. For both fully informative and uninformative priors the statistical significance of a shift in $\hat{\theta}^{A+B} - \hat{\theta}^A$ is the same as the shift in $\hat{\theta}^A - \hat{\theta}^B$ since both of them are weighted by their inverse covariance. We note that in non-update form and for uninformative priors, i.e. Eq. (4.3), parameter differences are equivalent to the Index of Inconsistency (Lin & Ishak, 2017b,a, 2019), while providing a clear assessment of statistical significance rather than interpretation on the Jeffreys' scale.

There are two main advantages of using $Q_{\rm UDM}$ instead of non-update difference in mean statistics: parameter-space directions that can exhibit interesting tension are identified *a priori*, i.e. before explicitly measuring the tension, to aid physical interpretation; non-Gaussianities are mitigated since we can select the most constraining and Gaussian of two data sets.

As shown in Raveri & Hu (2019), an effective method to compute $Q_{\rm UDM}$ in practice consists of breaking down the calculation as a sum over the Karhunen–Loéve (KL) modes of the covariances involved. We indicate these modes with ϕ^a and their corresponding generalized eigenvalue with λ^a. The modes ϕ^a are uncorrelated for both data set A and $A + B$. For a given KL mode $\lambda^a - 1$ is the improvement observed for the variance in the value of that mode when the second data set is added to the first. To avoid sampling noise in the calculation of $Q_{\rm UDM}$ we restrict our calculation to modes that satisfy:

$$0.2 < \lambda^a - 1 < 100\,. \tag{4.17}$$

The lower bound removes directions along which data set B is not updating A, while the upper bound removes directions along which A is not updating B. In both cases, with perfect knowledge of the covariances these directions would not contribute to the end result.

We notice here that the procedure of identifying the KL modes can be performed *a priori*, before looking at the data, starting from the Fisher matrix. We also point out that the set of KL modes is invariant under linear parameter transformations while the principal-component decomposition is not.

The KL decomposition of parameter shifts allows to investigate the physical origin of the reported tensions. As discussed in Wu et al. (2020) we can write the parameters' Fisher matrix $F = (C)^{-1}$ as a sum over KL components:

$$F_{\alpha\alpha} = \sum_a F^a_{\alpha\alpha} = \sum_a \phi^a_\alpha \phi^a_\alpha / \lambda^a\,. \tag{4.18}$$

The fractional Fisher information $F^a_{\alpha\alpha}/F_{\alpha\alpha} \in [0, 1]$ tells us how important a given KL mode is in constraining a cosmological parameter. Low values mean that the KL mode can be removed from the full decomposition without altering the parameter constraint.

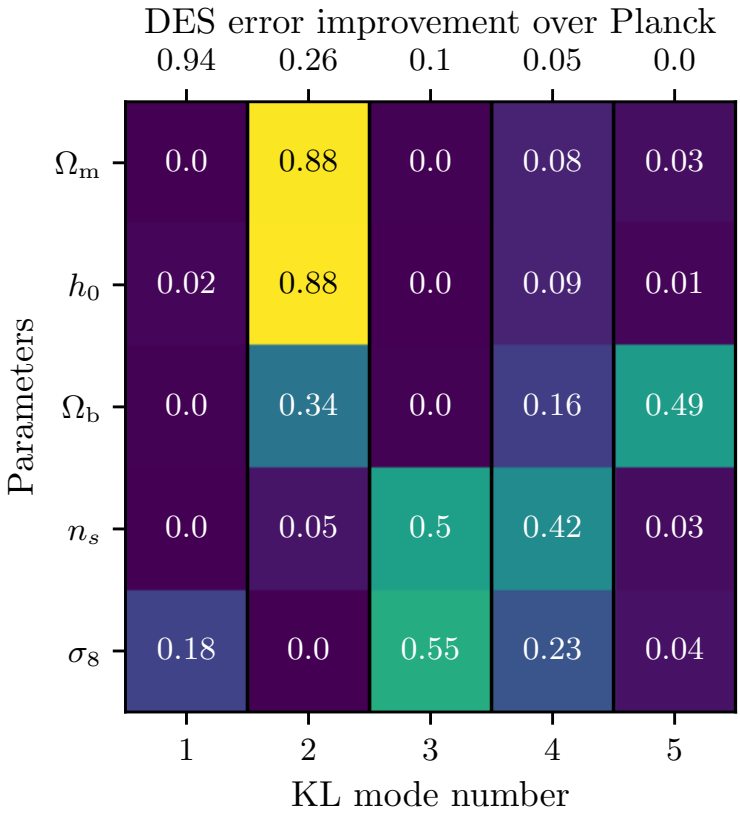

Figure 4.4.2: The fractional Fisher information on cosmological parameters for *Planck* computed using the KL modes from its update with simulated DES. Each line shows the fractional contribution of each KL mode to the total information on a given parameter. The sum of values in each row is one. The numbers on top of the figure show the fractional error improvement of DES over *Planck* for each KL mode.

In Fig. 4.4.2 we show the fractional contribution of different KL modes to the *Planck* Fisher matrix when it is updated with our simulated DES measurements. We also report in the figure the error improvement which is given by $\sqrt{\lambda^a} - 1$ for each mode. We have a total of five modes, equal to the number of parameters that the data sets have in common and we have sorted them by error improvement of DES+*Planck* over *Planck* alone. The first data set – in this case *Planck* – is setting the parameter combinations that are updated for each mode, while the second data set is setting the improvement factor. For the first two modes we can see that DES improves on the *Planck* determination of σ_8 by almost a factor two (94%) and the determination of $\Omega_{\rm m}h^2$ by 26%. DES does not improve other modes significantly.

We use the implementation of $Q_{\rm UDM}$ and related KL decomposition algorithms in the `tensiometer` code.

4.4.5 Goodness-of-fit loss

We next consider Goodness-of-fit loss which measures how much goodness-of-fit degrades when joining two data sets. This is a method in between evidence- and parameter-based ones since it relies on both likelihood values and parameters. When fitting two data sets separately, each probe can individually invest all model parameters in improving its goodness of fit. However, when the two measurements are joined the parameters have to compromise and the quality of the joint fit naturally degrades. This degradation is quantified by the estimator:

$$Q_{\rm DMAP} = 2\ln\mathcal{L}_A(\theta_{pA}) + 2\ln\mathcal{L}_B(\theta_{pB}) - 2\ln\mathcal{L}_{A+B}(\theta_{pA+B}) \tag{4.19}$$

where θ_{pA}, θ_{pB} and θ_{pA+B} are the Maximum a Posteriori (MAP) parameters measured by the first and second probe and their combination respectively, and $\mathcal{L}$ is the data likelihood for the single and joint probes and is evaluated at the Maximum a Posteriori (MAP) point, θ_p. We use the subscript DMAP to denote the difference in MAP estimates. As discussed in Raveri & Hu (2019), when the likelihoods and posteriors are Gaussian $Q_{\rm DMAP}$ is χ^2 distributed with

$$\Delta N_{\rm eff} = N^A_{\rm eff} + N^B_{\rm eff} - N^{A+B}_{\rm eff} \tag{4.20}$$

degrees of freedom where $N^A_{\rm eff}$, $N^B_{\rm eff}$, and $N^{A+B}_{\rm eff}$ are the respective numbers of the degrees of freedom

$$N_{\rm eff} = N - {\rm tr}[C_\Pi^{-1}C_p] \tag{4.21}$$

is the number of parameters that a data set ends up constraining compared to the priors it began with. The goodness-of-fit is expected to degrade by one for each measured parameter, and indicates tension if the decrease is higher. Only the parameters that are constrained by the data over the prior can contribute to a tension since prior-constrained parameters cannot be optimized to improve the data fit. In the limits where the prior is uninformative or fully informative $Q_{\rm DMAP}$ is the likelihood expression for parameter shifts discussed in the previous sections and its statistical significance should match the one obtained with parameter-shift techniques.

Notice that this estimator requires Gaussianity in both data space and parameter space. This is a stronger requirement than just approximate Gaussianity in parameter space, and limits its

applicability in practice. Most of the likelihoods that we use here are Gaussian in data space with the exception of the large-scale CMB likelihood. This can be thought to be a prior on the optical depth of reionization, τ, that would not contribute to the tension budget since it is not shared with DES and hence allows us to use $Q_{\rm DMAP}$.

We use the implementation of $Q_{\rm DMAP}$ in the `tensiometer` code.

4.4.6 Eigentension

The goal of the Eigentension parameter-space method is to identify well-measured eigenmodes in the data and compare the parameter constraints of two experiments within the subspace spanned by the well-measured eigenmodes. Here, we briefly describe the steps taken to quantify the tension between the fiducial *Planck* and DES constraints in this paper, and refer the reader to Park & Rozo (2019) for a more detailed discussion and testing of the method.

We begin by identifying the well-measured parameter subspace by following these steps:

1. Obtain the parameter covariance matrix from a set of fiducial constraints for DES and identify the eigenvectors of this covariance matrix.

2. For each eigenvector, take the ratio of its variance in the prior to its variance in the posterior. If this ratio is above 10^2, identify the eigenvector as well-measured or robust.

3. Project the fiducial *Planck* constraints and the various DES constraints along the subspace spanned by the robust eigenvector(s), and create importance sampled chains of equal length for each constraint.

For (i), we use constraints from a fiducial DES analysis with a noiseless data vector generated from theory under the *Planck* best-fit parameters and the true DES Y1 covariance matrix. This allows the *ad hoc* choice of 10^2 as the threshold value in (ii), which we make after examining the eigenvectors from (i), to be *a priori*. We identify one well-measured DES eigenvector:

$$e_{\rm DES} = \sigma_8 \Omega_{\rm m}^{0.57} \tag{4.22}$$

that has a variance ratio of 2665, and construct importance sampled chains of length 10^5 along this eigenmode. With the projected chains in hand, we quantify tension between two constraints i and j as following; we

1. construct the chain of differences $\Delta e = e_i - e_j$ between the importance sampled chains for i and j.

2. approximate the probability surface for Δe via KDE , and identify the iso-probability contour that crosses the origin, i.e. $\Delta e = 0^N$, where N is the number of robust eigenvectors identified.

3. integrate the probability surface within the origin-crossing contour, and convert the integral to Gaussian sigmas.

1D shift	a-priori	Bayes ratio		Eigentension	GoF Loss	MCMC/Update	Suspiciousness
	Tension	$\log R$	Interpretation			Param Diffs	
Baseline	$0\,\sigma$	5.7 ± 0.6	Strong Agreement	$0.5\,\sigma$	$0.2\,\sigma$	$0.3/0.3\,\sigma$	$(0.1 \pm 0.1)\,\sigma$
$\Delta\sigma_8 = -0.5 \times \delta\sigma_8$	$0.0\,\sigma$	6.4 ± 0.6	Strong Agreement	$0.4\,\sigma$	$0.4\,\sigma$	$0.3/0.4\,\sigma$	$(0.2 \pm 0.2)\,\sigma$
$\Delta\Omega_m = 0.5 \times \delta\Omega_m$	$0.1\,\sigma$	5.4 ± 0.6	Strong Agreement	$1.3\,\sigma$	$0.7\,\sigma$	$0.9/0.8\,\sigma$	$(0.5 \pm 0.2)\,\sigma$
$\Delta\sigma_8 = -1 \times \delta\sigma_8$	$0.4\,\sigma$	5.5 ± 0.6	Strong Agreement	$1.1\,\sigma$	$0.8\,\sigma$	$1.0/0.8\,\sigma$	$(0.3 \pm 0.2)\,\sigma$
$\Delta\Omega_m = 1 \times \delta\Omega_m$	$1.0\,\sigma$	3.5 ± 0.5	Strong Agreement	$2.3\,\sigma$	$1.9\,\sigma$	$1.8/1.7\,\sigma$	$(1.5 \pm 0.3)\,\sigma$
$\Delta\sigma_8 = -1.5 \times \delta\sigma_8$	$1.1\,\sigma$	3.6 ± 0.6	Strong Agreement	$2.0\,\sigma$	$1.2\,\sigma$	$1.8/1.9\,\sigma$	$(1.5 \pm 0.3)\,\sigma$
$\Delta\Omega_m = 1.5 \times \delta\Omega_m$	$2.3\,\sigma$	-0.4 ± 0.6	No Evidence	$3.3\,\sigma$	$3.0\,\sigma$	$2.8/2.7\,\sigma$	$(2.9 \pm 0.4)\,\sigma$
$\Delta\sigma_8 = -2 \times \delta\sigma_8$	$2.0\,\sigma$	0.3 ± 0.6	No Evidence	$2.6\,\sigma$	$2.1\,\sigma$	$2.7/3.0\,\sigma$	$(2.2 \pm 0.4)\,\sigma$
$\Delta\Omega_m = 2 \times \delta\Omega_m$	$3.8\,\sigma$	-4.8 ± 0.6	Strong Tension	$4.1\,\sigma$	$3.9\,\sigma$	$3.4/3.6\,\sigma$	$(4.1 \pm 0.6)\,\sigma$
$\Delta\sigma_8 = -3 \times \delta\sigma_8$	$3.7\,\sigma$	-6.2 ± 0.6	Strong Tension	$4.3\,\sigma$	$3.4\,\sigma$	$4.6/4.8\,\sigma$	$(3.7 \pm 0.5)\,\sigma$
$\Delta\Omega_m = 3 \times \delta\Omega_m$	$> 5\,\sigma$	-16.2 ± 0.6	Strong Tension	$> 5.4\,\sigma$	$6.2\,\sigma$	$5.3/5.3\,\sigma$	$(5.9 \pm 0.7)\,\sigma$
$\Delta\sigma_8 = -5 \times \delta\sigma_8$	$> 5\,\sigma$	-26.3 ± 0.6	Strong Tension	$> 5.4\,\sigma$	$5.8\,\sigma$	$6.8/8.8\,\sigma$	$(6.3 \pm 0.8)\,\sigma$
$\Delta\Omega_m = 5 \times \delta\Omega_m$	$> 5\,\sigma$	-47.0 ± 0.6	Strong Tension	$> 5.4\,\sigma$	$10.0\,\sigma$	$6.6/8.1\,\sigma$	$(9.6 \pm 1.2)\,\sigma$

Table 4.4.2: The tension between *Planck* and simulated DES chains for different shifts in σ_8 and Ω_m, calculated via the different tension metrics described in the main text. The first column refers to the number of one-dimensional standard deviations by which each parameter is shifted, defined in Eq. (4.2). The a-priori Gaussian tension is calculated as described in Sec. 4.3 and serves only as an order of magnitude approximation of expected results. The probability results of each of the tension metrics is converted to a number of effective sigmas using Eq. (4.4).

For (ii), we use a Gaussian KDE with bandwidths determined from Silverman's rule of thumb, and a straightforward Monte Carlo integration with 1.28×10^7 random draws,which is sufficient to quantify tensions up to 5.4σ.

4.4.7 Other metrics

As mentioned in the introductions, a plethora of methods to quantify tension can be found in the cosmological literature. Our work does not investigate all of these methods, as this would make the analysis too wide in scope. For example, Hyperparameters (Hobson et al., 2002; Luis Bernal & Peacock, 2018) are more useful to construct a posterior from data sets in tension, by factoring in possible unknown systematic effects. The surprise (Seehars et al., 2016) is best suited for experiments that are an update from a previous version with less data. Posterior Predictive Distributions (Feeney et al., 2019) are similar in nature to the Evidence Ratio as shown in Lemos et al. (2020). Other methods are not considered as they closely resemble others, such as Amendola et al. (2013); Martin et al. (2014); Joudaki et al. (2017) being based on the Bayesian Evidence ratio, and Lin & Ishak (2017a); Adhikari & Huterer (2019); Lin & Ishak (2019) being different versions of parameter differences in update form.

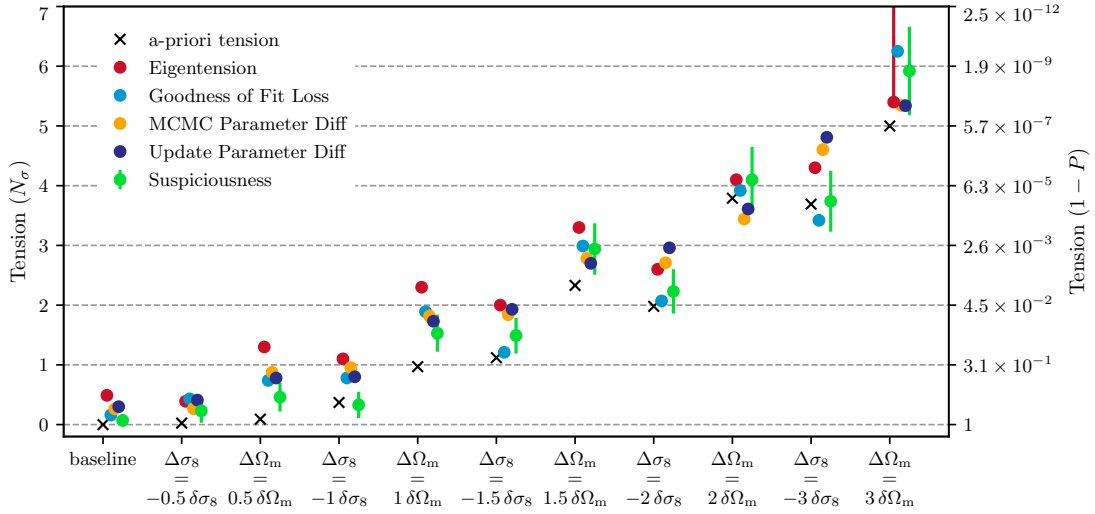

Figure 4.5.1: A graphical illustration of the main results of Tab. 4.4.2. Different points show the tension calculated by each tension metric as a function of the input shifts. The error bars in the green points correspond to sampling errors, which can be calculated for evidence-based methods by re-sampling the nested sampling weights.

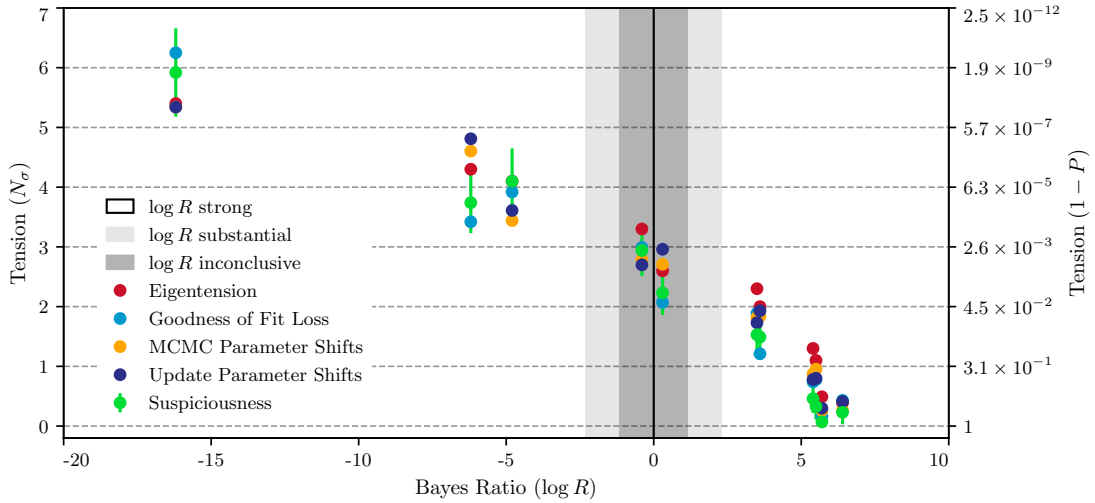

Figure 4.5.2: Tension estimates given by different metrics versus the corresponding Bayes ratio. Shaded regions highlight Jeffreys' scale used to interpret the Bayes ratio, with the vertical line separating "Tension" to the left and "Agreement" to the right.

4.5 Results using simulated DES data

In this section, we apply the tension metrics described in Sec. 4.4 to the simulated vectors obtained as outlined in Sec. 4.3, and compare the results to our a-priori expectation from Sec. 4.3. Our results are shown in Tab. 4.4.2 and graphically illustrated in Fig. 4.5.1.

We first note that our estimates of a-priori Gaussian tension should be only used as an rough indication and are generally lower than the tension evaluated by the metrics that we study. This is because the a-priori Gaussian tension does not have noise in the data vector while the tensions simulations do. This noise realization is the same for all the shifts, which explains the fact that the a-priori tension is systematically lower in all results with respect to other tension estimators. We can see this in the baseline case, where in a noiseless case all metrics would obtain perfect agreement (a '0σ' tension), but instead the noise leads to small discrepancies.

When applying parameter-shift estimators in both MCMC and update form we can see, from Tab. 4.4.2 and Fig. 4.5.1, that, for tensions measured up to 5σ, the two estimates agree very well, to within 0.3σ. This overall result is reassuring since these two estimators are measuring the same sense of tension between the two data sets. This agreement is also expected since the distributions that we consider are roughly Gaussian in the bulk of the distribution. At high statistical significance MCMC results are lower in both cases and this suggests that the decay of the tails of the distribution is slower than a Gaussian distribution. For the parameter update we observe that the two parameter combinations, discussed in Sec. 4.4.4, that DES+*Planck* significantly improves over *Planck*-only do not appreciably change throughout the test cases.

In case of either fully informative or uninformative priors the statistical significance of GoF loss is expected to match the one reported by parameter-shift estimators. As we can see from Tab. 4.4.2 that is the case at low statistical significance. Non-Gaussianities in the form of slowly-decaying tails violate the assumptions used by the GoF loss estimator, while their impact can be mitigated by parameter shifts in update form. As a result, as statistical significance increases, in Tab. 4.4.2 the two estimates differ. In particular, as expected, GoF loss overestimates statistical significance since this estimator is assuming Gaussian decay in the tails.

For Eigentension, we make use of the metric on the simulated vectors, making use of the robust DES eigenvector and the Monte Carlo sampling procedure discussed in Sec. 4.4.6. Note that the Eigentension metrics are calculated only up to 5.4σ, or 1 in 1.28×10^7; beyond this probability we simply quote that the tension is greater than 5.4σ and consider the tension to be definitive. The results are in good agreement with other tension metrics, in particular the two parameter shift estimators, with which Eigentension shares the general approach of quantifying tensions at the parameter space level.

With Suspiciousness, as shown in Tab. 4.4.2 and in Fig. 4.5.1, we obtain good agreement with the rest of tension metrics, especially when we consider the sampling error estimated from repeated re-samplings for the weights of the chain. To assign a tension probability, we need to calculate the Bayesian Model Dimensionality, for which we get $d = 2.3 \pm 0.1$. At high statistical significance, Suspiciousness seems to agree particularly well with GoF loss. This is reassuring since the two estimators coincide in the Gaussian limit with uninformative priors.

In Tab. 4.4.2 we also show the results for the Bayes ratio, interpreted with the Jeffreys' scale as used by DES Collaboration (2018), and shown in Tab. 4.4.1. As we can see from the table the interpretation of R transitions very quickly from 'Strong Agreement' to 'Strong Tension'. To further investigate the relation between R and the other metrics we plot them against each other in Fig. 4.5.2. This immediately highlights that the Jeffreys' scale that we use to interpret the Bayes ratio results lacks granularity in how it quantifies physical tensions. Coherently across different

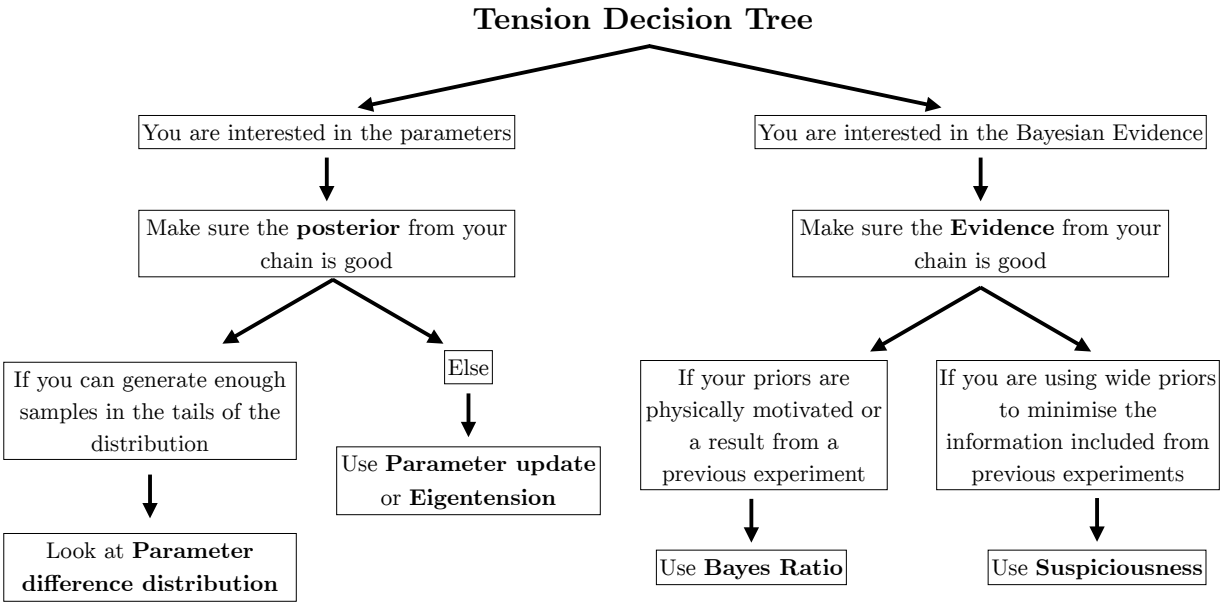

Figure 4.5.3: A practical 'decision tree' to measure tension, illustrating when each tension metric should be used.

estimators the interpretation of R goes from one extreme case to the other in a probability interval that covers about one standard deviation. Fig. 4.5.2 also clearly shows the bias of the evidence ratio toward agreement. The value of $R = 1$, which separates agreement and disagreement for our choice of priors is at a probability level that roughly corresponds to 3σ (i.e. a probability of the discrepancy occurring by chance of $p_T \sim 0.003$). We note that the offset between $R = 1$ and 50% probability events is set by the prior width and would hence change when changing the prior. Fig. 4.5.2 also shows that the evidence ratio, interpreted with the Jeffreys' scale, would still signal a strong tension, if present, while lacking granularity in the discrimination of mildly statistically significant tensions.

In Sec. 4.4, we made a distinction between parameter-space methods and evidence-based methods. We find that all our tension metrics agree well not only amongst themselves, but also qualitatively with the a-priori Gaussian tension calculations described in Sec. 4.3. This is a non-trivial result, as both the calculations and the fundamental questions that the various methods are trying to address differ.

The only exceptions to this good agreement are given by the statistically-significant σ_8 shifts where the spread between the three parameter difference estimators is smaller than the difference between them GoF loss and Suspiciousness; and the smaller a-priori shifts in Ω_m, for which the a-priori Gaussian tension estimate is smaller than the results from Eigentension and Suspiciousness. Since the input calculation used a noiseless data vector and simulated DES data vectors had noise,

these disagreements are expected. They are likely to be caused by the noise introduced in the chains used by the tension metrics, and will have a more significant impact on the small shifts.

Based on these results, we propose a methodology to quantify tension between data sets that exploits the strengths of all the different methods, summarized by Fig. 4.5.3. Within the parameter-based approach, we recommend to generate a **Monte-Carlo parameter difference distribution** and observe where the zero-difference point stands provided we have enough samples of the posterior distribution in its tail, as this method has no problem with non-Gaussianities, and has the advantage of providing useful visualizations in the form of confidence regions generated directly from the difference chain itself. However, if the number of samples in the tension tail is insufficient, this parameter-difference distribution will not be reliable enough to make statements about tension. In this case, either **Eigentension** or **parameter differences in update form** provide reliable metrics of tension. These two methods are also useful in identifying the physics behind the tension, as they provide characteristic parameter combinations along with the identified tensions lie. Since it does not offer mitigation of non-Gaussianities, we do not recommend using goodness-of-fit loss on its own, but rather as a cross-check with other metrics.

For the Evidence-based methods, if we have a well-motivated prior, such as the posterior from a previous experiment or a physically-motivated one, we can calculate the tension using the **Bayes ratio**. However, as discussed in the text, experiments such as DES and *Planck* often choose wide priors in order to obtain posteriors that do not depend on previous experiments. The arbitrariness in the choice of width of those priors means that we cannot use the Bayes ratio, as discussed in Sec. 4.4.1, unless we calibrated R using Fig. 4.5.2, but that would require recalibration if any details of the analysis changed. In the case of wide and uninformative priors, the **Suspiciousness** answers the same question as the Bayes ratio but correcting for the prior volume effect. We recommend its use over the Bayes ratio in general since it has the additional desirable property of having a 'tension probability' interpretation under a Gaussian approximation, without any need for calibration.

As pointed out in Fig. 4.5.3, different methods requires reliable calculations of different quantities. Parameter-space methods require a good estimate of the posterior, and particularly of its mean and covariance matrix. Evidence-based methods require a calculation of the Bayesian Evidence. Therefore, our choice of tension metric should inform our sampling choices, as further discussed in Lemos et al. (2022).

4.6 Application to DES Y1 and Planck

With a better understanding of the interpretation of each of the tension metrics, we now revisit the issue of consistency between the DES Y1 cosmology results and those obtained by the *Planck* collaboration (Planck Collaboration et al., 2016, 2018). This also serves as a worked example on real data of how tension between experiments can be fully quantified.

We choose to investigate three different combinations of DES data sets: (1) weak lensing-only constraints from Troxel et al. (2018) (2) constraints from combining the auto and cross-correlation between weak lensing and galaxy clustering, referred to as the 3×2pt analysis: (3) constraints from

data set	Bayes ratio		Eigentension	GoF Loss	MCMC/Update	Suspiciousness
	log R	Interpretation			Param Shifts	
DES cosmic shear vs. *Planck* 15	2.2 ± 0.5	Substantial Agreement	$1.8\,\sigma$	$1.3\,\sigma$	$1.3/1.2\,\sigma$	$(0.7 \pm 0.4)\,\sigma$
DES 3×2pt vs. *Planck* 15	1.0 ± 0.5	No Evidence	$2.4\,\sigma$	$2.7\,\sigma$	$2.2/2.2\,\sigma$	$(2.4 \pm 0.2)\,\sigma$
DES 5×2pt vs. *Planck* 15	1.1 ± 0.5	Substantial Agreement	$2.4\,\sigma$	$2.8\,\sigma$	$2.1/2.3\,\sigma$	$(2.2 \pm 0.3)\,\sigma$
DES 5×2pt vs. *Planck* 15 + lensing	1.0 ± 0.6	No Evidence	$2.4\,\sigma$	$2.5\,\sigma$	$2.1/2.3\,\sigma$	$(2.2 \pm 0.4)\,\sigma$
DES 5×2pt + *Planck* lensing vs. *Planck* 15	6.1 ± 0.6	Strong Agreement	$1.6\,\sigma$	$2.4\,\sigma$	$1.9/2.2\,\sigma$	$(1.8 \pm 0.2)\,\sigma$
DES cosmic shear vs. *Planck* 18	3.3 ± 0.4	Strong Agreement	$1.5\,\sigma$	$1.0\,\sigma$	$1.0/1.1\,\sigma$	$(0.5 \pm 0.3)\,\sigma$
DES 3×2pt vs. *Planck* 18	2.2 ± 0.6	Substantial Agreement	$2.2\,\sigma$	$1.6\,\sigma$	$2.0/2.3\,\sigma$	$(2.4 \pm 0.2)\,\sigma$

Table 4.6.1: The tension between *Planck* and different data set combinations involving DES Y1 data, calculated via the different tension metrics described in the main text. In the first column, *Planck* refers to the combination of the TT, TE and EE likelihoods. In bold font we highlight the combinations of DES 3×2pt and *Planck*, as those are the main focus of this section. The horizontal line separates *Planck* 2015 and 2018 data set combinations.

(2) plus cross-correlation with CMB lensing, referred as the 5×2pt analysis (DES Collaboration, 2019a). We particularly focus in the second combination, as it provided the most powerful constraints from large-scale structure measured by DES alone. For *Planck* 2015 we use the small-scale ($\ell > 30$) measurements of the CMB temperature power spectrum and the joint large-scale temperature and polarization data. For *Planck* 2018 we use small-scale CMB temperature, polarization and their cross-correlation measurements combined with large-scale temperature and and E-mode polarization data. In doing so we follow the recommendations of the *Planck* collaboration in the two data releases.

The results of parameter estimation for these data sets are shown in Fig. 4.6.1 and the results of different tension estimators in Tab. 4.6.1. We highlight in the table the results that we focus our discussion on.[9]

We start with MCMC parameter shifts, as it is the parameter-based method that can give the most accurate value for the tension, thanks to its ability to go beyond the Gaussian approximation. In Fig. 4.6.2 we can see the posterior of differences between the determination of σ_8 and Ω_m from different DES data sets and *Planck* that clearly shows a tension that is greater than 2σ. In Tab. 4.6.1 we see that in full parameter space this tension is at the 2.2σ level. We proceed with Suspiciousness as our recommended evidence-based method which fully confirms the parameter-shift results, giving a $2.4 \pm 0.2\sigma$ tension between *Planck* 2015 and DES 3×2pt. We note that applying both methods provides a useful cross-check of their respective results. This moderate tension remains when *Planck* is updated from the 2015 to the 2018 data and for DES 5×2pt. This shows that this tension is robust to the inclusion of CMB polarization data.

To understand the physics behind these discrepancies, it is useful to consider other methods. Using Eigentension, we identify a single well-measured eigenmode for each DES analysis: $\sigma_8\Omega_\mathrm{m}^{0.57}$

[9]The reader might notice that the values of the Bayes ratio reported in Tab. 4.6.1, in particular for the case DES 3×2pt vs. *Planck* 15, differ from the values reported by DES Collaboration (2018) ($R = 6.6$). This difference has been identified as originating from sampling issues in the DES Y1 analysis, as will be described in more detail in Lemos et al. (2022).

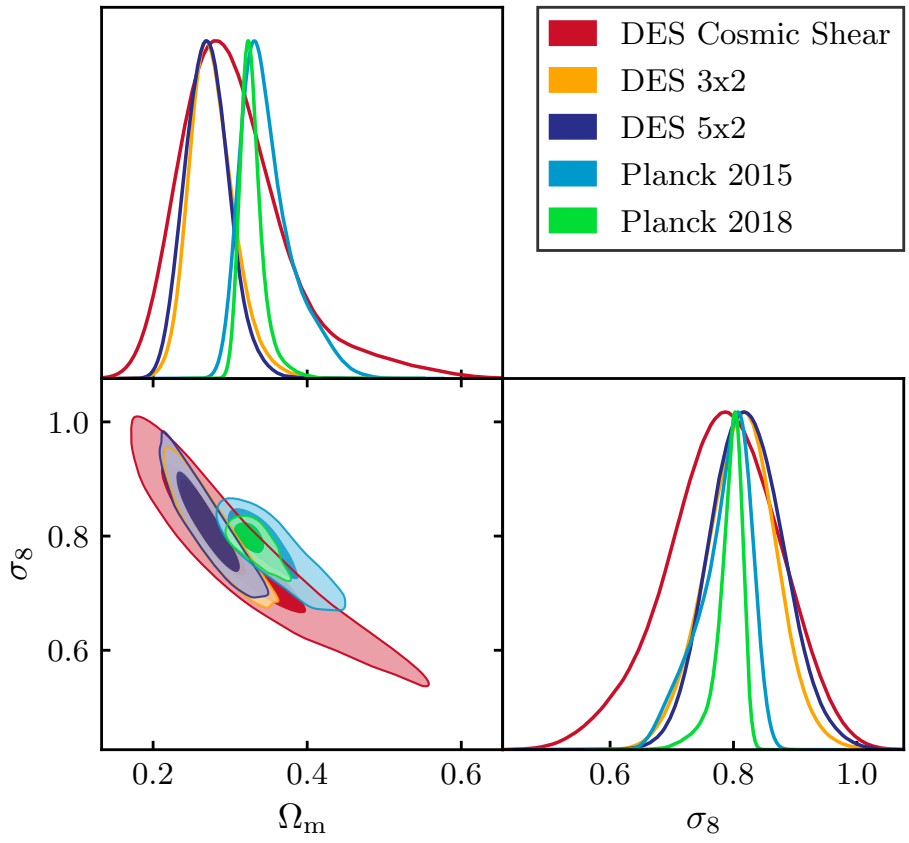

Figure 4.6.1: 68% and 95% confidence regions of the joint marginalized posterior probability distributions for Dark Energy Survey Year 1 Cosmic Shear, 3×2pt and 5×2pt likelihoods, and for the *Planck* 2015 TTTEEE likelihood.

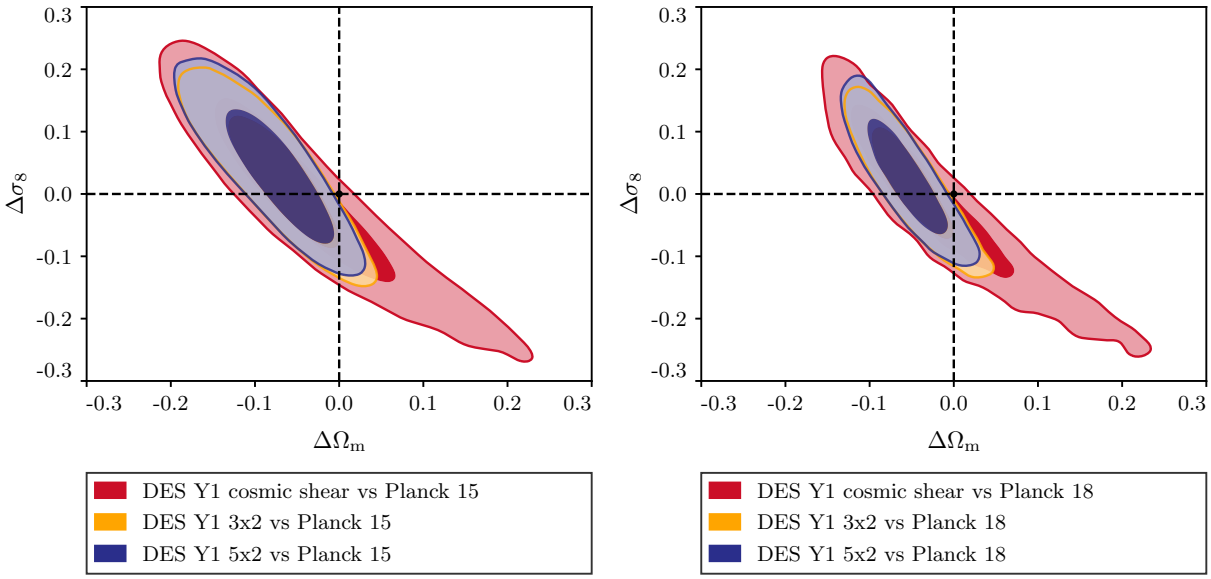

Figure 4.6.2: Joint marginalized posterior distribution of the parameter differences between different DES data selections and *Planck* 15/18. The distribution of parameter differences is used to compute the statistical significance of a parameter shift. The darker and lighter shading corresponds to the 68% and 95% C.L. regions respectively.

for the 3×2pt analysis, and $\sigma_8 \Omega_{\rm m}^{0.58}$ in the 5×2pt case. Both eigenmodes are very similar to the widely-used definition of $S_8 = \sigma_8 (\Omega_{\rm m}/0.3)^{0.5}$, and can be interpreted as representing the 'lensing strength' arising from the large-scale structure of the late-time Universe. After measuring tension exclusively along this direction in parameter space, we find results that are in agreement with other methods. This shows that the moderate tension between DES and *Planck* is found along a parameter space direction that we believe DES is robustly measuring. Studying parameter updates of DES with respect to *Planck* gives similar conclusions. As discussed in the previous section and shown in Fig. 4.4.2, combining DES improves the *Planck* determination of two parameters, the first mode projecting mostly onto σ_8 and the second onto $\Omega_{\rm m} h^2$. The first mode drives most of the tension while the shift in the second is compatible with a statistical fluctuation. Decrease in Goodness of Fit agrees with other estimators.

The Bayes ratio interpreted on the Jeffreys' scale reports no significant tension between all data combinations that we consider. Given the results of the previous section we can understand this as the data tension not overcoming the bias of the Bayes ratio toward agreement. We note that the priors used for the fiducial analyses in the previous section do not coincide with the priors used in this section; we thus cannot use the previously-derived calibration of the Bayes ratio.

The mild tension we obtain between *Planck* and DES, varying between 2σ and 3σ, should not be overlooked. While this level of tension could still be a statistical fluke, it is significant enough to warrant in-depth future investigations. The forthcoming DES Y3 analysis, incorporating a larger fraction of the sky, is expected to shed light on this matter.

4.7 Conclusions

In this work, we have explored different methods to quantify consistency between two uncorrelated data sets, focusing on the comparison between DES and *Planck*. The motivation is to decide on a metric of tension between these two surveys ahead of the DES Y3 data release. This was done by simulating a set of DES data sets with values of cosmological parameters chosen to introduce varying levels of discrepancy with *Planck*. We calculate the tension for each simulated DES data set, and compare to an a-priori Gaussian tension expected based on the known true cosmologies for the simulated data sets. While this work has been performed for the specific case of DES and *Planck*, our findings about the different metrics described in Sec. 4.5 apply to any problem of tension quantification. However, if we wanted to apply the Bayes ratio to a different problem with uninformative priors, the exercise of calibrating the Bayes ratio would have to be repeated.

We have found that the Bayes' ratio used in the Y1 analysis has several flaws that make it unsuitable for the quantitative comparison of DES and *Planck*. In particular, it is proportional to the width of the chosen uninformative prior; it relies on the Jeffreys' scale to interpret the ratio of probabilities, which needs an unknown calibration that is problem-dependent (i.e. we would need to build a table such as Tab. 4.4.2 in every problem to calculate the overall calibration of the Bayes ratio); and the fact that we can only calculate logarithms of the probability ratio means that the Jeffreys' scale used in the DES Y1 analysis (Tab. 4.4.1) will in most cases diagnose extreme agreement or extreme tension.

As shown in Tab. 4.4.2, the other four tension metrics employed in this work — Eigentension, GoF loss, Parameter differences, and Suspiciousness — agree with the a-priori tension, as well as amongst themselves, with the exceptions of small shifts in Ω_m and large shifts in σ_8 discussed in Sec. 4.5, which are likely the result of noise introduced in the simulated data vectors. We conclude that any of the tension metrics can be used for the problem of quantifying tension between DES and *Planck*, as they produce similar results.

We use these tension metrics to re-assess the tension between DES Y1 and *Planck* 2015, as well as with the latest *Planck* 2018 results. We find, similar to our findings from the simulated analyses, that the dependence of the Evidence ratio on calibration causes the results to be inconsistent with what we see in the plots, and what all other tension metrics indicate. We find that there is a $\sim 2.3\sigma$ between DES and *Planck*, which remains when the *Planck* 2018 likelihood is used. It remains to be seen how this will evolve when the more powerful DES Y3 data are used. If the tension is reduced when more data are considered, we are likely looking at a statistical fluctuation. If the tension remains or increases, we could be looking at unexplained systematics in either of the surveys, or evidence of physics beyond the ΛCDM model.

Data availability Statement

The data underlying this article are available in the Dark Energy Survey Data Management platform, at `https://des.ncsa.illinois.edu`

Acknowledgements

Funding for the DES Projects has been provided by the U.S. Department of Energy, the U.S. National Science Foundation, the Ministry of Science and Education of Spain, the Science and Technology Facilities Council of the United Kingdom, the Higher Education Funding Council for England, the National Center for Supercomputing Applications at the University of Illinois at Urbana-Champaign, the Kavli Institute of Cosmological Physics at the University of Chicago, the Center for Cosmology and Astro-Particle Physics at the Ohio State University, the Mitchell Institute for Fundamental Physics and Astronomy at Texas A&M University, Financiadora de Estudos e Projetos, Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro, Conselho Nacional de Desenvolvimento Científico e Tecnológico and the Ministério da Ciência, Tecnologia e Inovação, the Deutsche Forschungsgemeinschaft and the Collaborating Institutions in the Dark Energy Survey.

The Collaborating Institutions are Argonne National Laboratory, the University of California at Santa Cruz, the University of Cambridge, Centro de Investigaciones Energéticas, Medioambientales y Tecnológicas-Madrid, the University of Chicago, University College London, the DES-Brazil Consortium, the University of Edinburgh, the Eidgenössische Technische Hochschule (ETH) Zürich, Fermi National Accelerator Laboratory, the University of Illinois at Urbana-Champaign, the Institut de Ciències de l'Espai (IEEC/CSIC), the Institut de Física d'Altes Energies, Lawrence Berkeley National Laboratory, the Ludwig-Maximilians Universität München

and the associated Excellence Cluster Universe, the University of Michigan, NFS's NOIRLab, the University of Nottingham, The Ohio State University, the University of Pennsylvania, the University of Portsmouth, SLAC National Accelerator Laboratory, Stanford University, the University of Sussex, Texas A&M University, and the OzDES Membership Consortium.

Based in part on observations at Cerro Tololo Inter-American Observatory at NSF's NOIRLab (NOIRLab Prop. ID 2012B-0001; PI: J. Frieman), which is managed by the Association of Universities for Research in Astronomy (AURA) under a cooperative agreement with the National Science Foundation.

The DES data management system is supported by the National Science Foundation under Grant Numbers AST-1138766 and AST-1536171. The DES participants from Spanish institutions are partially supported by MICINN under grants ESP2017-89838, PGC2018-094773, PGC2018-102021, SEV-2016-0588, SEV-2016-0597, and MDM-2015-0509, some of which include ERDF funds from the European Union. IFAE is partially funded by the CERCA program of the Generalitat de Catalunya. Research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) including ERC grant agreements 240672, 291329, and 306478. We acknowledge support from the Brazilian Instituto Nacional de Ciência e Tecnologia (INCT) do e-Universo (CNPq grant 465376/2014-2).

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

PL acknowledges STFC Consolidated Grants ST/R000476/1 and ST/T000473/1. We also thank the organizers of the DES Y3 workshop: "Probing Dark Energy Observations in the Nonlinear Regime" at the University of Michigan in Ann Arbor, where this project started.

Appendix

4.A Dark Energy Survey data

The Dark Energy Survey (DES, DES Collaboration, 2005, 2016b) is a six-year survey that has observed over 5000 deg^2 in five filters ($grizY$) and has probed redshifts up to $z \sim 1.3$. It has also used time-domain to measure several thousand type Ia supernovae (SNe Ia). DES can constrain cosmological parameters in several ways: It can use these SNe Ia, and treat them as standarizable candles to constrain cosmology through their redshift-luminosity relation, usually referred to as Hubble Diagram (Hubble, 1929; Kirshner, 2004); it can use the distribution of galaxies to measure the Baryon Acoustic Oscillation (BAO) feature which was imprinted by sound waves at the recombination era ($z \sim 1100$), and which serves as a standard ruler (Eisenstein et al., 2007); it can use the abundance of galaxy clusters, the largest gravitationally-bound structures in the Universe (Allen et al., 2011); it can use the distribution of galaxies to measure the dark matter density distribution, under the assumption of some bias relating the two, called galaxy clustering; and it can measure the distortion of light by intervening matter along the line of sight, referred to as gravitational lensing (Mandelbaum, 2018). When the matter distribution distorting the path of light is the large-scale structure of the Universe, the effect is called cosmic shear (Kilbinger, 2015). Because in this case distortions are too small to be detected for individual galaxies, they are detected through correlations in the shapes and position of galaxies images.

Using data from the first year of observations (Y1), the DES collaboration has already reported constraints on cosmology from BAO (DES Collaboration, 2019c), galaxy clustering (Elvin-Poole et al., 2018), cosmic shear (Troxel et al., 2018), the cross-correlation of galaxy clustering and cosmic shear, referred to as galaxy–galaxy lensing (Prat et al.,, 2018), and as a main result, the combination of the two-point functions from cosmic shear, galaxy clustering, and galaxy–galaxy lensing, henceforth referred to as '3×2pt' (DES Collaboration, 2018). In addition, using data from three years of observations (Y3), DES has also constrained cosmology from SNe Ia (DES Collaboration, 2019d), and galaxy clusters (To et al., 2020). However, as described in DES Collaboration (2019b), the most powerful constraints from future DES data releases will come from combinations of the different probes, as these can break degeneracies in parameter constraints and significantly increase accuracy.

We adopt the same priors used in the DES Y1 analysis, shown in Tab. 4.A.1.

Parameter	Prior
Cosmology	
$\Omega_{\rm m}$	flat (0.1, 0.9)
$A_{\rm s}$	flat $(5 \times 10^{-10}, 5 \times 10^{-9})$
$n_{\rm s}$	flat (0.87, 1.07)
$\Omega_{\rm b}$	flat (0.03, 0.07)
h	flat (0.55, 0.90)
$\Omega_\nu h^2$	flat$(5 \times 10^{-4}, 10^{-2})$
Lens Galaxy Bias	
$b_i (i = 1, 5)$	flat (0.8, 3.0)
Intrinsic Alignment	
$A_{\rm IA}$	flat $(-5, 5)$
$\eta_{\rm IA}$	flat $(-5, 5)$
Lens photo-z shift (red sequence)	
$\Delta z_{\rm l}^1$	Gauss (0.0, 0.007)
$\Delta z_{\rm l}^2$	Gauss (0.0, 0.007)
$\Delta z_{\rm l}^3$	Gauss (0.0, 0.006)
$\Delta z_{\rm l}^4$	Gauss (0.0, 0.01)
$\Delta z_{\rm l}^5$	Gauss (0.0, 0.01)
Source photo-z shift	
$\Delta z_{\rm s}^1$	Gauss (0.0, 0.016)
$\Delta z_{\rm s}^2$	Gauss (0.0, 0.013)
$\Delta z_{\rm s}^3$	Gauss (0.0, 0.011)
$\Delta z_{\rm s}^4$	Gauss (0.0, 0.022)
Shear calibration	
$m^i (i = 1, 4)$	Gauss (0.0, 0.023)

Table 4.A.1: Cosmological and nuisance parameters and their priors used in this analysis.

Chapter 5

Conclusions

The work presented in this thesis delineates significant advancements in the domain of observational cosmology, facilitated by the integration of artificial intelligence methodologies, novel statistical methods, and extensive galactic data sets. The methodologies developed are not only being utilized to interpret existing data from ongoing surveys but also hold promise for advancing the analysis of future surveys.

Chapter 2 introduces a refined Self-Organizing Map (SOM) algorithm, the SOMPZ method, which has notably enhanced the precision of photometric redshift estimation. Its successful application to the Dark Energy Survey's Y3 data has dramatically minimized redshift bin overlaps, enriching our understanding of weak lensing phenomena. Such progress not only elevates the DES Year 6 analysis but also lays a solid groundwork for upcoming cosmic studies. The SOMPZ Y6 pipeline I developed is now a public resource, and is being integrated into the RAIL ecosystem (Schmidt et al., 2023), which will facilitate the characterization of redshift distributions in several contexts, but specially for the case of LSST.

Despite these advancements, however, photometric redshift uncertainty stands as a complex challenge for the coming decade's wide-field imaging surveys. There is an urgent need for in-depth research to devise strategies for comparing and synthesizing various photometric redshift techniques against the backdrop of expansive, yet incomplete, spectroscopic redshift data sets. As we venture into the era of observing billions of galaxies, the rich spectroscopic data from instruments like the Dark Energy Spectroscopic Instrument and the Roman HLS prism survey are invaluable. Such data sets are pivotal for enhancing redshift calibration, especially at higher redshifts and fainter magnitudes, solidifying spectroscopic data's role as the cornerstone of this methodology.

The empirical methodology for model selection introduced in Chapter 3 encapsulates a nuanced balance between parameter bias mitigation and model complexity. This approach, predicated upon the calibration of synthetic data, provides a robust framework for interpreting χ^2 discrepancies, thereby facilitating informed model selection in the absence of uninformative priors, while maintaining the data blinded. Applied to the intrinsic alignment challenge within weak lensing surveys, this methodology permits a more nuanced comparison of the NLA and TATT models, contributing to a refined understanding of which model is sufficient to describe

the data. One caveat of this method, however, is its dependency on the stability of other analysis components. To ensure that χ^2 differences are solely attributable to the model under test, this method should be applied as one of the final steps in the analysis process.

Furthermore, the implementation of tension estimators in assessing the congruence between DES and *Planck* measurements, in Chapter 4, shows that the existing tension metrics succeed in capturing existing tensions in the entire multi-dimensional parameter space. This work also offers guidance on how to compute these metrics in a reliable way, combining the information provided by the evidence and by parameter space. It revealed a discernible tension at the level of 2.3σ within the ΛCDM paradigm for DES Y1 and Planck 2018, which subsequently got reduced to 0.7σ in DES Y3 (DES Collaboration, 2022). This points towards the difference being due to statistical fluctuations. If that is the case, as the precision of future surveys increase, we should keep seeing a increasing in agreement. If this trend is indeed statistical, the precision of forthcoming surveys should reveal an increasing concordance. However, considering that these tension metrics were validated using a DES Y1 setup, it is crucial to reassess their effectiveness with future data sets from Rubin, Roman, and Euclid surveys to ensure they still reliably detect tensions.

In conclusion, this research has expanded the horizons of cosmological data analysis, yet it represents merely one step in a much larger journey of discovery. Standing at the threshold of a revolutionary phase in cosmology, I am optimistic that the methodologies and insights derived from this work will shed some light on the path forward. The influx of data from cutting-edge facilities, particularly the Vera C. Rubin Observatory, will not only test the robustness of the methods developed but also open avenues for their refinement. As we venture into this new era, it is crucial that the approaches outlined here undergo rigorous further testing and refinement. The apparent tensions between various cosmological observations call for a more exhaustive examination. The need for adaptation of these methods to a broader range of cosmological probes is also clear. Moving forward, the challenge will be to not only refine our current models but also to innovate new frameworks that could provide a more holistic understanding of the universe. This thesis lays the groundwork for some of the transformative science that awaits us in the forthcoming epoch of next-generation surveys.

Bibliography

Adhikari S., Huterer D., 2019, JCAP, 1901, 036

Aihara H., et al., 2018, Publications of the ASJ, 70, S8

Akaike H., 1973, Proceedings of the Second International Symposium on Information Theory, Tsahkadsor, Armenia, USSR

Albrecht A., Steinhardt P. J., 1982, Phys. Rev. Lett., 48, 1220

Allen S. W., Evrard A. E., Mantz A. B., 2011, Annual Review of Astron and Astrophysis, 49, 409

Alpher R. A., Herman R., Gamow G. A., 1948, Phys. Rev., 74, 1198

Amendola L., Marra V., Quartin M., 2013, Mon. Not. Roy. Astron. Soc., 430, 1867

Amon A., et al., 2022, Phys. Rev. D, 105, 023514

Andrae R., Schulze-Hartung T., Melchior P., 2010, arXiv e-prints, p. arXiv:1012.3754

Asgari M., et al., 2020, arXiv e-prints, p. arXiv:2007.15633

Asgari M., et al., 2021, Astronomy and Astrophysics, 645, A104

Benítez N., 2000, Astrophysical Journal, 536, 571

Bennett C. L., et al., 1994, Astrophysical Journal, 436, 423

Bennett C. L., et al., 2003, The Astrophysical Journal Supplement Series, 148, 1

Bird S., Viel M., Haehnelt M. G., 2012, Monthly Notices of the Royal Astronomical Society, 420, 2551

Bishop C. M., 2006, Pattern Recognition and Machine Learning. Springer, Berlin, Heidelberg

Blake C., Bridle S., 2005, Monthly Notices of the Royal Astronomical Society, 363, 1329

Blazek J., Vlah Z., Seljak U., 2015, Journal of Cosmology and Astroparticle Physics, 8, 015

Blazek J. A., MacCrann N., Troxel M. A., Fang X., 2019, Physical Review D, 100, 103506

Box G. E. P., Cox D. R., 1964, Journal of the Royal Statistical Society. Series B (Methodological), 26, 211

Bridges M., Feroz F., Hobson M. P., Lasenby A. N., 2009, Monthly Notices of the Royal Astronomical Society, 400, 1075

Bridle S., King L., 2007, New Journal of Physics, 9, 444

Bridle S., et al., 2010, Monthly Notices of the Royal Astronomical Society, 405, 2044

Buchs R., et al., 2019, Mon. Not. Roy. Astron. Soc., 489, 820

Campos A., Samuroff S., Mandelbaum R., 2023, Monthly Notices of the Royal Astronomical Society, 525, 1885

Carrasco Kind M., Brunner R. J., 2013, Monthly Notices of the Royal Astronomical Society, 432, 1483

Carrasco Kind M., Brunner R. J., 2014, Monthly Notices of the Royal Astronomical Society, 438, 3409

Catelan P., Kamionkowski M., Blandford R. D., 2001, Monthly Notices of the Royal Astronomical Society, 320, L7

Chacón J., Duong T., 2018, Multivariate Kernel Smoothing and Its Applications. Chapman & Hall/CRC Monographs on Statistics and Applied Probability, CRC Press

Charnock T., Battye R. A., Moss A., 2017, Phys. Rev., D95, 123535

Chen A., et al., 2023, Monthly Notices of the Royal Astronomical Society, 518, 5340

Colless M., et al., 2001, Monthly Notices of the Royal Astronomical Society, 328, 1039

Collister A. A., Lahav O., 2004, Publications of the ASP, 116, 345

Crittenden R. G., Natarajan P., Pen U.-L., Theuns T., 2002, Astrophysical Journal, 568, 20

DES Collaboration 2005, arXiv e-prints, pp astro–ph/0510346

DES Collaboration 2016a, Physical Review D, 94, 022001

DES Collaboration 2016b, Mon. Not. Roy. Astron. Soc., 460, 1270

DES Collaboration 2018, Physical Review D, 98, 043526

DES Collaboration 2019a, Phys. Rev. D, 100, 023541

DES Collaboration 2019b, doi:10.1103/PhysRevLett.122.171301, 122, 171301

DES Collaboration 2019c, Monthly Notices of the Royal Astronomical Society, 483, 4866

DES Collaboration 2019d, Astrophysical Journal, 872, L30

DES Collaboration 2022, Physical Review D, 105, 023520

DESI Collaboration 2023, The Early Data Release of the Dark Energy Spectroscopic Instrument, doi:10.5281/ZENODO.7964161, https://zenodo.org/record/7964161

Dalal R., et al., 2023, Hyper Suprime-Cam Year 3 Results: Cosmology from Cosmic Shear Power Spectra (arXiv:2304.00701)

Dark Energy Survey Collaboration Kilo-Degree Survey Collaboration 2023, The Open Journal of Astrophysics, 6, 36

Dawson K. S., et al., 2013, Astronomical Journal, 145, 10

DeRose J., et al., 2019, Astrophysical Journal, 875, 69

Desjacques V., Jeong D., Schmidt F., 2018, Physics Reports, 733, 1

Dodelson S., 2017, Gravitational lensing. Cambridge University Press, Cambridge, UK

Dodelson S., Schmidt F., 2020, Modern Cosmology, second edition. edn. Academic Press, London; United Kingdom

Doux C., Baxter E., et al., 2021, Monthly Notices of the Royal Astronomical Society, 503, 2688

Doux C., et al., 2022, Monthly Notices of the Royal Astronomical Society,

Efstathiou G., Sutherland W. J., Maddox S. J., 1990, Nature, 348, 705

Einstein A., 1916, Annalen der Physik, 354, 769

Eisenstein D. J., Seo H.-J., White M., 2007, Astrophysical Journal, 664, 660

Elvin-Poole J., et al., 2018, Physical Review D, 98, 042006

Euclid Collaboration et al., 2020, A&A, 644, A31

Everett S., et al., 2022, The Astrophysical Journal Supplement Series, 258, 15

Fang X., Blazek J. A., McEwen J. E., Hirata C. M., 2017, Journal of Cosmology and Astroparticle Physics, 2017, 030

Fang X., Eifler T., Krause E., 2020, Monthly Notices of the Royal Astronomical Society, 497, 2699

Feeney S. M., Peiris H. V., Williamson A. R., Nissanke S. M., Mortlock D. J., Alsing J., Scolnic D., 2019, Phys. Rev. Lett., 122, 061105

Feroz F., Hobson M. P., Bridges M., 2009, Monthly Notices of the Royal Astronomical Society, 398, 1601

Feroz F., Hobson M. P., Cameron E., Pettitt A. N., 2019, The Open Journal of Astrophysics, 2, 10

Fosalba P., Crocce M., Gaztañaga E., Castander F. J., 2015, Monthly Notices of the Royal Astronomical Society, 448, 2987

Friedman A., 1922, Zeitschrift für Physik, 10, 377

Friedrich O., et al., 2021, Monthly Notices of the Royal Astronomical Society, 508, 3125

Gamow G., 1946, Phys. Rev., 70, 572

Gatti M., et al., 2021, Mon. Not. Roy. Astron. Soc., 504, 4312

Geller M. J., Huchra J. P., 1989, Science, 246, 897

Gelman A., Carlin J. B., Stern H. S., Rubin D. B., 2004, Bayesian data analysis; 2nd ed.. Chapman and Hall, Boca Raton, FL, `https://cds.cern.ch/record/1010408`

Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press

Grandis S., Seehars S., Refregier A., Amara A., Nicola A., 2016, Journal of Cosmology and Astro-Particle Physics, 2016, 034

Guth A. H., 1981, Phys. Rev. D, 23, 347

Hamana T., et al., 2020, Publications of the ASJ, 72, 16

Handley W., 2019, The Journal of Open Source Software, 4, 1414

Handley W., Lemos P., 2019a, Physical Review D, 100, 023512

Handley W., Lemos P., 2019b, Phys. Rev., D100, 043504

Handley W. J., Hobson M. P., Lasenby A. N., 2015a, Monthly Notices of the Royal Astronomical Society, 450, L61

Handley W. J., Hobson M. P., Lasenby A. N., 2015b, Monthly Notices of the Royal Astronomical Society, 453, 4384

Hartley W. G., et al., 2021, Monthly Notices of the Royal Astronomical Society, 509, 3547

Heymans C., et al., 2006, Monthly Notices of the Royal Astronomical Society, 368, 1323

Heymans C., et al., 2013, Monthly Notices of the Royal Astronomical Society, 432, 2433

Heymans C., et al., 2020, arXiv e-prints, p. arXiv:2007.15632

Higson E., Handley W., Hobson M., Lasenby A., 2018, Bayesian Analysis, 13, 873

Hikage C., et al., 2019, Publications of the ASJ, 71, 43

Hildebrandt H., et al., 2017, Monthly Notices of the Royal Astronomical Society, 465, 1454

Hirata C. M., Seljak U., 2004, Physical Review D, 70, 063526

Hirata C. M., Seljak U. c. v., 2010, Phys. Rev. D, 82, 049901

Hirata C. M., Mandelbaum R., Ishak M., Seljak U., Nichol R., Pimbblet K. A., Ross N. P., Wake D., 2007, Monthly Notices of the Royal Astronomical Society, 381, 1197

Hobson M. P., Bridle S. L., Lahav O., 2002, Mon. Not. Roy. Astron. Soc., 335, 377

Hosoya A., Buchert T., Morita M., 2004, Physical Review Letters, 92, 141302

Hu W., 2000, Phys. Rev. D, 62, 043007

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

Ilbert O., et al., 2009, Astrophysical Journal, 690, 1236

Ivezić Ž., et al., 2019, Astrophysical Journal, 873, 111

Jarvis M., et al., 2021, Monthly Notices of the Royal Astronomical Society, 501, 1282

Jee M. J., Tyson J. A., Hilbert S., Schneider M. D., Schmidt S., Wittman D., 2016, Astrophysical Journal, 824, 77

Jeffreys H., 1961, The theory of probability. OUP Oxford

Jeffreys H., 1998, The Theory of Probability. Oxford Classic Texts in the Physical Sciences, OUP Oxford, https://books.google.se/books?id=vh9Act9rtzQC

Joachimi B., Taylor A. N., 2011, Monthly Notices of the Royal Astronomical Society, 416, 1010

Joachimi B., et al., 2015, Space Science Reviews, 193, 1

Joachimi B., et al., 2021a, Astronomy and Astrophysics, 646, A129

Joachimi B., Köhlinger F., Handley W., Lemos P., 2021b, Astronomy and Astrophysics, 647, L5

Johnston H., et al., 2019, Astronomy and Astrophysics, 624, A30

Joudaki S., et al., 2017, Mon. Not. Roy. Astron. Soc., 465, 2033

Joudaki S., Ferreira P. G., Lima N. A., Winther H. A., 2022, Phys. Rev. D, 105, 043522

Kaiser N., 1984, Astrophysical Journal, Letters, 284, L9

Kass R., Raftery A., 1995, Journal of the American Statistical Association, 90, 773

Kerscher M., Weller J., 2019, SciPost Phys. Lect. Notes, p. 9

Kiessling A., et al., 2015, Space Science Reviews, 193, 67

Kilbinger M., 2015, Reports on Progress in Physics, 78, 086901

Kilbinger M., et al., 2010, Monthly Notices of the Royal Astronomical Society, 405, 2381

Kirk D., et al., 2015, Space Science Reviews, 193, 139

Kirshner R. P., 2004, Proceedings of the National Academy of Science, 101, 8

Knabenhans M., et al., 2021, Monthly Notices of the Royal Astronomical Society, 505, 2840

Kohonen T., 1982, Biological Cybernetics, 43, 59

Krause E., Eifler T., 2017, Monthly Notices of the Royal Astronomical Society, 470, 2100

Krause E., et al., 2017, arXiv e-prints, p. arXiv:1706.09359

Krause E., et al., 2021, arXiv e-prints, p. arXiv:2105.13548

Krauss L. M., Turner M. S., 1995, Gen. Rel. Grav., 27, 1137

Kullback S., Leibler R. A., 1951, Ann. Math. Statist., 22, 79

Kunz M., Trotta R., Parkinson D. R., 2006, Physical Review D, 74, 023503

Laureijs R., et al., 2011, arXiv e-prints, p. arXiv:1110.3193

Lemaître G., 1927, Annales de la Société scientifique de Bruxelles, 47, 49

Lemos P., Köhlinger F., Handley W., Joachimi B., Whiteway L., Lahav O., 2020, Mon. Not. Roy. Astron. Soc., 496, 4647

Lemos P., Raveri M., Campos A., et al., 2021, Monthly Notices of the Royal Astronomical Society, 505, 6179

Lemos P., Weaverdyck N., et al., 2022, Monthly Notices of the Royal Astronomical Society,

Levi M., et al., 2013, The DESI Experiment, a whitepaper for Snowmass 2013 (arXiv:1308.0847)

Lewis A., 2019, arXiv e-prints, p. arXiv:1910.13970

Lewis A., Bridle S., 2002, Physical Review D, 66, 103511

Lewis A., Challinor A., Lasenby A., 2000, Astrophysical Journal, 538, 473

Li X., et al., 2023, Hyper Suprime-Cam Year 3 Results: Cosmology from Cosmic Shear Two-point Correlation Functions (arXiv:2304.00702)

Liddle A. R., 2007, Monthly Notices of the Royal Astronomical Society, 377, L74

Liddle A., 2015, An introduction to modern cosmology, third edition, 3rd ed. edn. Wiley

Liddle A., Mukherjee P., Parkinson D., 2006a, Astronomy & Geophysics, 47, 4.30

Liddle A. R., Mukherjee P., Parkinson D., Wang Y., 2006b, Physical Review D, 74, 123506

Limber D. N., 1953, Astrophysical Journal, 117, 134

Lin W., Ishak M., 2017a, Phys. Rev., D96, 023532

Lin W., Ishak M., 2017b, Phys. Rev., D96, 083532

Lin W., Ishak M., 2019, arXiv:1909.10991

Linde A., 1982, Physics Letters B, 108, 389

LoVerde M., Afshordi N., 2008, Phys. Rev. D, 78, 123506

Loureiro A., et al., 2022, Astronomy and Astrophysics, 665, A56

Luis Bernal J., Peacock J. A., 2018, JCAP, 1807, 002

Mandelbaum R., 2018, Annual Review of Astron and Astrophysis, 56, 393

Mandelbaum R., et al., 2008, Monthly Notices of the Royal Astronomical Society, 386, 781

Mandelbaum R., et al., 2015, Monthly Notices of the Royal Astronomical Society, 450, 2963

Marshall P., Rajguru N., Slosar A., 2006, Physical Review D, 73, 067302

Martin J., Ringeval C., Trotta R., Vennin V., 2014, Phys. Rev., D90, 063501

Masters D., et al., 2015, Astrophysical Journal, 813, 53

McEwen J. E., Fang X., Hirata C. M., Blazek J. A., 2016, Journal of Cosmology and Astroparticle Physics, 2016, 015

Mead A. J., Brieden S., Tröster T., Heymans C., 2021, Monthly Notices of the Royal Astronomical Society, 502, 1401

Miranda V., Rogozenski P., Krause E., 2020, arXiv:2009.14241

Miyatake H., et al., 2023, Hyper Suprime-Cam Year 3 Results: Cosmology from Galaxy Clustering and Weak Lensing with HSC and SDSS using the Emulator Based Halo Model (arXiv:2304.00704)

Myles J., Alarcon A., et al., 2021, Monthly Notices of the Royal Astronomical Society, 505, 4249

Neal R. M., 1998, arXiv e-prints, p. physics/9803008

Nicola A., Amara A., Refregier A., 2019, Journal of Cosmology and Astro-Particle Physics, 2019, 011

Osato K., Shirasaki M., Yoshida N., 2015, Astrophysical Journal, 806, 186

Ostriker J. P., Steinhardt P. J., 1995, Nature, 377, 600

Padilla L. E., Tellez L. O., Escamilla L. A., Vazquez J. A., 2019, arXiv e-prints, p. arXiv:1903.11127

Pandey S., et al., 2020, Physical Review D, 102, 123522

Park Y., Rozo E., 2019, arXiv e-prints, p. arXiv:1907.05798

Peebles P. J. E., 1966, Phys. Rev. Lett., 16, 410

Peebles P. J. E., 1984, Astrophysical Journal, 284, 439

Penzias A. A., Wilson R. W., 1965, Astrophysical Journal, 142, 419

Perlmutter S., et al., 1999, Astrophys. J., 517, 565

Planck Collaboration et al., 2016, Astronomy and Astrophysics, 594, A13

Planck Collaboration et al., 2018, preprint, (arXiv:1807.06209)

Prat J., et al., 2018, Physical Review D, 98, 042005

Raveri M., Hu W., 2019, Physical Review D, 99, 043506

Raveri M., Zacharegkas G., Hu W., 2020, Phys. Rev. D, 101, 103527

Riess A. G., et al., 1998, Astron. J., 116, 1009

Riess A. G., Casertano S., Yuan W., Macri L. M., Scolnic D., 2019, Astrophys. J., 876, 85

Rigdon E. E., 1999, Structural Equation Modeling: A Multidisciplinary Journal, 6, 219

Robertson H. P., 1929, Proceedings of the National Academy of Science, 15, 822

Rubin V. C., Ford W. Kent J., 1970, Astrophysical Journal, 159, 379

Saito S., Takada M., Taruya A., 2008, Physical Review Letters, 100, 191301

Samuroff S., et al., 2019, Monthly Notices of the Royal Astronomical Society, 489, 5453

Sánchez C., Raveri M., Alarcon A., Bernstein G. M., 2020, Monthly Notices of the Royal Astronomical Society, 498, 2984

Sánchez C., et al., 2022a, arXiv e-prints, p. arXiv:2211.16593

Sánchez C., Prat J., et al., 2022b, Physical Review D, 105, 083529

Schermelleh-Engel K., Moosbrugger H., Müller H., 2003, Methods of Psychological Research Online, 8, 23–74

Schmidt S., et al., 2023, LSSTDESC/RAIL: v0.98.5, doi:10.5281/zenodo.7927358, https://doi.org/10.5281/zenodo.7927358

Schneider P., van Waerbeke L., Mellier Y., 2002, Astronomy and Astrophysics, 389, 729

Schuhmann R. L., Joachimi B., Peiris H. V., 2016, Monthly Notices of the Royal Astronomical Society, 459, 1916

Schwarz G., 1978, Annals of Statistics, 6, 461–464

Secco L. F., Samuroff S., et al., 2022, Physical Review D, 105, 023515

Seehars S., Amara A., Refregier A., Paranjape A., Akeret J., 2014, Physical Review D, 90, 023533

Seehars S., Grandis S., Amara A., Refregier A., 2016, Physical Review D, 93, 103507

Simon P., Hilbert S., 2018, Astronomy and Astrophysics, 613, A15

Singh S., Mandelbaum R., 2016, Monthly Notices of the Royal Astronomical Society, 457, 2301

Skilling J., 2006, Bayesian Anal., 1, 833

Slipher V. M., 1917, Proceedings of the American Philosophical Society, 56, 403

Speagle J. S., et al., 2019, Monthly Notices of the Royal Astronomical Society, 490, 5658

Spergel D., et al., 2015, arXiv e-prints, p. arXiv:1503.03757

Spiegelhalter D. J., Best N. G., Carlin B. P., van der Linde A., 2001, Bayesian Measures of Model Complexity and Fit

Spiegelhalter D. J., Best N. G., Carlin B. P., Van Der Linde A., 2002, Journal of the Royal Statistical Society Series B, 64, 583

Steiger J., Shapiro A., Browne M., 1985, Psychometrika, 50, 253

Suchyta E., et al., 2016, Monthly Notices of the Royal Astronomical Society, 457, 786

Sugiyama S., et al., 2023, Hyper Suprime-Cam Year 3 Results: Cosmology from Galaxy Clustering and Weak Lensing with HSC and SDSS using the Minimal Bias Model (arXiv:2304.00705)

Takahashi R., Sato M., Nishimichi T., Taruya A., Oguri M., 2012, Astrophysical Journal, 761, 152

The LSST Dark Energy Science Collaboration et al., 2018, The LSST Dark Energy Science Collaboration (DESC) Science Requirements Document (arXiv:1809.01669), doi:10.48550/arXiv.1809.01669

To C., et al., 2020, arXiv e-prints, p. arXiv:2010.01138

Tokdar S. T., Kass R. E., 2010, WIREs Computational Statistics, 2, 54

Tröster T., et al., 2022, Astronomy and Astrophysics, 660, A27

Trotta R., 2007, Monthly Notices of the Royal Astronomical Society, 378, 72

Trotta R., 2008, Contemporary Physics, 49, 71

Troxel M. A., Ishak M., 2015, Physics Reports, 558, 1

Troxel M. A., et al., 2018, Physical Review D, 98, 043528

Tyson J. A., Valdes F., Jarvis J. F., Mills A. P. J., 1984, Astrophysical Journal, Letters, 281, L59

Valentino E. D., et al., 2021, Classical and Quantum Gravity, 38, 153001

Vardanyan M., Trotta R., Silk J., 2011, Monthly Notices of the Royal Astronomical Society, 413, L91

Verde L., Protopapas P., Jimenez R., 2013, Physics of the Dark Universe, 2, 166

Verde L., Treu T., Riess A. G., 2019, Nature Astronomy, 3, 891

Walker A. G., 1937, Proceedings of the London Mathematical Society, 42, 90

Weaverdyck N., Alves O., et al., 2022, in prep

Wilks S. S., 1938, The Annals of Mathematical Statistics, 9, 60

Wright A. H., Hildebrandt H., van den Busch J. L., Heymans C., 2020, Astronomy & Astrophysics, 637, A100

Wu W. K., Motloch P., Hu W., Raveri M., 2020, Phys. Rev. D, 102, 023510

Yang J., Turner M. S., Steigman G., Schramm D. N., Olive K. A., 1984, Astrophysical Journal, 281, 493

York D. G., et al., 2000, The Astronomical Journal, 120, 1579

Zuntz J., et al., 2015, Astronomy and Computing, 12, 45

de Jong J. T. A., Verdoes Kleijn G. A., Kuijken K. H., Valentijn E. A., KiDS Consortiums A.-W., 2013, Experimental Astronomy, 35, 25