Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **534,** 2142–2167 (2024) https://doi.org/10.1093/mnras/stae2082
Advance Access publication 2024 September 9

Separating states in astronomical sources using hidden Markov models: with a case study of flaring and quiescence on EV Lac

Robert Zimmerman,[1] David A. van Dyk,[2★] Vinay L. Kashyap[3★] and Aneta Siemiginowska[3]

[1] *Department of Statistical Sciences, University of Toronto, 700 University Avenue, Toronto, ON M5G 1Z5 , Canada*
[2] *Statistics Section, Department of Mathematics, Imperial College London, 180 Queen's Gate, London, SW7 2AZ, UK*
[3] *Center for Astrophysics | Harvard and Smithsonian, 60 Garden Street, Cambridge, MA 02138 , USA*

Accepted 2024 August 30. Received 2024 August 22; in original form 2024 May 10

ABSTRACT

We present a new method to distinguish between different states (e.g. high and low, quiescent and flaring) in astronomical sources with count data. The method models the underlying physical process as latent variables following a continuous-space Markov chain that determines the expected Poisson counts in observed light curves in multiple passbands. For the underlying state process, we consider several autoregressive processes, yielding continuous-space hidden Markov models of varying complexity. Under these models, we can infer the state that the object is in at any given time. The continuous state predictions from these models are then dichotomized with the help of a finite mixture model to produce state classifications. We apply these techniques to X-ray data from the active dMe flare star EV Lac, splitting the data into quiescent and flaring states. We find that a first-order vector autoregressive process efficiently separates flaring from quiescence: flaring occurs over 30 per cent–40 per cent of the observation durations, a well-defined persistent quiescent state can be identified, and the flaring state is characterized by higher plasma temperatures and emission measures.

Key words: methods: data analysis – methods: statistical – stars: coronae – stars: flare – stars: individual: EV Lac – X-rays: stars.

1 INTRODUCTION

The ubiquitous variability of astronomical sources spans large dynamic ranges in both intensity and time-scale. The intensities typically vary differently in different passbands (i.e. they exhibit spectral variations as well). The causes of such variability are diverse, ranging from nuclear flashes occurring in low-mass X-ray binaries over durations of seconds, to magnetic reconnection flares on stars and accretion-driven dipping in compact binaries lasting from a fraction of a ks to tens of ks, to gravitational lensing lasting for days, to abrupt changes in accretion levels onto compact objects which then persist for long durations ranging from weeks to months, to cyclic activity on stars that spans a decade, etc. The underlying physical processes that lead to such strong variations are not fully understood. In order to model and predict these variations, we first need to identify robustly the times when the states of the sources appear to change.

We posit here that when we observe large intermittent variability, there is some identifiable characteristic in the source system – modelled as a hidden state – which serves as a predictor to distinguish between different levels of activity. As an example, consider the flaring activity on stars, where we observe short duration bursts whose profiles show a rapid rise in intensity exceeding the typical intensity by several factors, followed by a cooling-dominated exponential decay. This profile manifests as a stochastic sequence of alternating active periods with frequent and energetic emissions at short time-scales of a few ks, and quiescent periods with periodic or smaller fluctuations. We aim to build a model that describes the timing of such flaring, and includes a rudimentary quantification of the underlying variability. From a statistical point of view, including a latent process enables us to model observed correlations in the light curve and thus to predict and estimate the long-run proportion of time spent in flaring and quiescent states.

Previous work on detecting or isolating such variability has focused mainly on local statistical significance testing, applying a set of somewhat ad-hoc rules, using automatic/black-box learning methods (e.g. neural networks) to identify flares in observed light curves, or modelling the intensities as a mixture distribution. In a study of γ-ray flares in blazars, for example, Nalewajko (2013) used a simple rule that first identifies the peak flux and then defines the flare duration as the time interval with flux greater than 50 per cent of that observed in the peak. Robinson et al. (1995) took a more statistical approach in their search for microflares in dMe flare stars: they computed the statistical significance of peaks in the binned data where the null distribution is determined by repeating their procedure on light curves where the bins have been randomly permuted. Aschwanden & Freeland (2012) proposed an 'automated flare detection algorithm' which is a set of criteria that are applied to a smoothed light curve; a background/quiescent level is determined using the time period before a local minimum in the light curve and the flare is associated with the interval starting at this minimum and continuing through the first subsequent local minimum that is

★ E-mail: d.van-dyk@imperial.ac.uk (DAvD); vkashyap@cfa.harvard.edu (VLK) below a background-dependent threshold. Peck et al. (2021) adopted a similar procedure to detect flares in Geostationary Operational Environmental Satellite (GOES) X-ray light curves.[1] A large sample of M-dwarf flares was obtained by Davenport et al. (2014) using an iterative smoothing procedure to remove star-spot and then identifying flares as intervals that exhibit a positive flux excursion of more than 2.5σ. More recently, supervised learning methods such as convolutional neural networks (e.g. Feinstein et al. 2020) have been used, while other researchers have continued to rely on visual inspection (e.g. Kashapova et al. 2021). Nearly all efforts to date have focused on univariate single-band light curves. A notable exception appears in Fleming et al. (2022), who combined near-UV (ultraviolet) and far-UV light curves in a search for flares in M-dwarfs. They deployed a set of rules whereby a (peak) flare is identified by either two consecutive NUV data points above 3σ or two simultaneous data points above 3σ, one in each band.

While these methods include techniques that make use of statistical significance and standard deviations, they do not take advantage of principled statistical methods to model or fit features in the observed light curves. More principled statistical methods for identifying 'bursts' in astrophysical light curves were pioneered by Scargle's work on Bayesian Blocks (Scargle 1998; Scargle et al. 2013). The method assumes a piecewise constant intensity function for a Poisson process in time, and implements a fully Bayesian strategy for estimating the number of breakpoints. The time intervals with constant intensity are called blocks and their number is determined by maximizing the Bayes factor or posterior odds. The breakpoints are determined sequentially via their posterior distribution as blocks are added to the model. The Bayesian Blocks method has proved to be an invaluable tool for identifying 'bursts' in light curves and has recently been used to separate the quiescent and active states of γ-ray flaring blazars (Yoshida et al. 2023). However, because the adopted model is piecewise constant, the fit results become difficult to interpret when dealing with smoothly increasing or decreasing intensities.

Large variability in astronomical sources is inevitably accompanied by spectral changes. In the case of stellar X-ray variability, Wong et al. (2016) proposed using a marked Poisson process for photon arrivals, treating photon wavelength as a 'mark'. As with Bayesian Blocks, their method, called Automark, assumes a piecewise constant intensity function for the Poisson process that governs photon arrivals. Spectra are assumed to be constant between the breakpoints, but within each block are modelled in a flexible non-parametric manner that accounts for spectral lines. The number of breakpoints is determined via the minimum description length principle. The method was extended to include spatial information/images by Xu et al. (2021).

Neither Bayesian Blocks nor Automark provides a mechanism to model the underlying processes that generate the flares. With solar data the observation of individual flares enables a set of different but also principled statistical approaches. Focusing exclusively on timing data for solar flares, for example, a number of authors have used characteristics of the distribution of waiting times between solar flares to better understand the process generating the flares. In this way, researchers have concluded that the waiting-time distribution is consistent with a time-varying Poisson process (e.g. Wheatland 2000; Moon et al. 2001; Wheatland & Litvinenko 2002; Aschwanden 2019) or have used it to study the memory in this underlying process (e.g. Lepreti, Carbone & Veltri 2001; Lei et al. 2020; Rivera et al. 2022). Unfortunately, these techniques do not apply to stars other than the Sun because individual flares are not observable.

In this paper, we consider the specific case of X-ray flares in stellar coronae, where we seek to model not the individual flares but rather the underlying flaring states, allowing us to estimate the flaring fraction and to study the spectra in different states. To this end, we employ a discrete-time hidden Markov model (HMM; Zucchini, MacDonald & Langrock 2017). This involves formulating a latent discrete-time Markov chain to represent the flaring process and is done in discrete time to match the discrete-time nature of the observed data. One novelty of our approach is that it leverages multiband light curves to identify flaring and quiescent intervals. The flaring process evolves as a Markov chain over time and in each time interval the chain's value determines the distribution of the observed counts, and thus influences the evolution of the observed data over time. We consider both the case where the latent flare process can enter one of a finite number of states (e.g. a quiescent state and an active state) and the case of a continuum of states through which the process evolves. Mathematically, these two possibilities correspond to discrete and continuous state spaces of the latent Markov chain.

We use two EV Lac light curves as a case study for our methods and find empirically that the continuous state-space HMM provides a better representation of the light curves than does the discrete-state-space HMM. However, the continuous-space HMM poses a computational challenge because its likelihood is intractable. Thus, we introduce an approximation that is based on a truncated and discretized state space and that can be made arbitrarily precise. We propose three specific formulations of the continuous-space HMM for flaring stars and a method for choosing among these formulation. We then fit the preferred model and use it to estimate the underlying continuous state variable that indexes the transition between the quiescent and active states. Below, we denote this (possibly multivariate) indexing variable as $\boldsymbol{X}_t$.

The continuous-space HMM does not clearly differentiate between the quiescent and active states of the source, instead allowing for variability within the states and a smooth transition between them. None the less, we aim to estimate the flaring fraction and to study the spectra within each state. As such, we introduce a two-state analysis, where Stage 1 fits a continuous-space HMM and estimates the continuous state indexing variable $\boldsymbol{X}_t$ and Stage 2 fits a finite mixture model to $\boldsymbol{X}_t$ in order to estimate the actual intervals of quiescence and activity. The Markov process underlying the HMM allows us to model the temporal autocorrelations evident in the light curves and thus to capture them in the fitted $\boldsymbol{X}_t$. In the second stage, we ignore these autocorrelations and focus instead on the marginal fitted values of $\boldsymbol{X}_t$ and use them to quantify the source's transitions between quiescence and activity. In this way, we can identify the long-run proportion of time spent in quiescence and flaring activity. The state predictions also allow us to estimate time intervals of quiescence and flaring, from which we obtain a comparative spectral analysis of both quiescence and flaring.

To the best of our knowledge, HMMs were first used to model time-series of flare data by Stanislavsky et al. (2020), who used a two-state autoregressive HMM to model continuous-valued daily solar X-ray flux emission data in an effort to study the hidden process underlying solar flares. They focused primarily on next-day prediction of solar flare activity. More recently, Esquivel et al. (2024) used a similar approach with three states to model the flaring activity of an M-dwarf star, in which the light curve was observed in one optical band with the *TESS* (*Transiting Exoplanet Survey Satellite*) Observatory.

[1] See also appendix A of the User's Guide for GOES-R XRS L2 Products by Machol, Codrescu, & Peck (data.ngdc.noaa.gov/platforms/solar-space-observing-satellites/goes/goes16/l2/docs/GOES-R_XRS_L2_Data_Users_Guide.pdf; July 2024).

HMMs have also been used in other applications in astrophysics, such as distinguishing between noise- and source-dominated states on strongly variable sources such as Sgr A* (Meyer et al. 2014).

Our approach here is more general. We use X-ray event lists containing information on photon arrival times and photon energy to construct light curves in multiple bands with low count rates in the Poisson regime, allowing us to explore short time-scale events as well as spectral variations. While our method allows for prediction, our primary aim is to better understand the underlying physical process driving stellar flares.

The remainder of this paper is organized into six sections. We begin by introducing two EV Lac light curves in Section 2 to motivate our modelling choice. Section 3 consists of a general introduction and review of HMMs, emphasizing the notation and properties needed in the current setting. We present our Stage 1 analysis with its three HMMs in Section 4, emphasizing techniques for quantifying uncertainty and model selection.We turn to the Stage 2 analysis in Section 5 with a new proposed model-based method for classifying light curves into flaring and quiescent intervals. We illustrate the application of these models and methods with an analysis of the EV Lac light curves in Section 6. Finally, we conclude with a discussion and suggestions for future work in Section 7. Several appendices review details of the algorithms used for maximum-likelihood fitting of discrete-space HMMs, present technical aspects of the discrete approximation that we use for efficient fitting of continuous-space HMMs, and give additional details of our analysis of EV Lac.

2 DATA

To motivate the development of HMMs as a modelling tool for non-periodic stochastic variability, we focus on stellar flares in particular, as those data sets often provide a clean look at a quiescent level punctuated by large, short-duration flares. Being able to separate quiescent from flaring states is crucial to understand mechanisms of stellar coronal heating, as well as the local interplanetary environment. The latter in particular affects the habitability of exoplanets, which has been flagged as an important focus of investigations in the Astro 2020 Decadal Survey (National Academies of Science, Engineering, and Medicine 2021).

2.1 EV Lac

The nearby (5 pc) active dMe binary EV Lac is a good candidate to test our HMM modelling. It has displayed consistent flaring across decades (at $\gtrsim 0.2 - 0.4$ h^{-1} during every X-ray observation; see Huenemoerder et al. 2010, and references therein), and there are high-spectral and high-temporal resolution, long-duration data sets obtained using the high-energy transmission gratings (Canizares et al. 2005) on the *Chandra X-ray Observatory* (Weisskopf et al. 2002).[2] These data were previously analysed by Huenemoerder et al. (2010), who detected 25 large individual flares across the data sets, and observed clear changes in spectral characteristics during flares, with generally higher plasma temperatures ($\gg 10^6$ K) at larger emission measures; they explicitly demonstrate the value of stacking the data from flares (whether short or long) and the quiescent durations.

Here, we use the combined dispersed events from both the high-energy (HEG) and medium-energy (MEG) grating components of the first-order photons, extracted from the level-2 event list using the default extraction regions in CIAO v4.16 (Fruscione et al. 2006). This allows us to avoid pileup effects (Davis 2001) on the zeroth-order data, especially during strong flares. We show the light curves for both epochs in Fig. 1, with the data split into two passbands, a softer band covering 0.3–1.5 keV and a harder band covering 1.5–8.0 keV. The choice of 1.5 keV as the split threshold is driven by the effective area peaking at that value.[3] There are approximately 23,600 and 17,900 counts in the softer band, and approximately 9,800 and 9,500 counts in the harder band for ObsIDs 01885 and 10679, respectively. The counts are collected into light curves (Fig. 1) binned at 50 s (see Appendix D for a sensitivity analysis for the choice of bins). Because these light curves are constructed from dispersed photons, pileup is entirely ignorable. The data are not affected by dead time effects, and background contamination is small and unvarying, and therefore also ignorable. The Advanced CCD Imaging Array - Spectroscopic detector on Chandra (ACIS-S) contamination build up at low energies over the mission (Plucinsky, Bogdan & Marshall 2022) reduces the counts in the soft band.

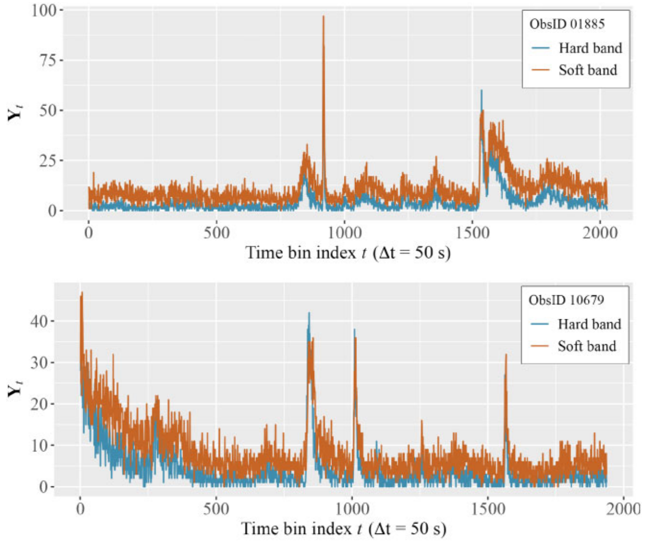

Figure 1. Bivariate time-series plots of EV Lac count data based on event lists where the split is based on counts in soft (0.3–1.5 keV) and hard (1.5–8 keV) passbands. Time is discretized into 50 s intervals; for ObsID 01855 (above), $t = 0$ corresponds to 2001 September 19, 19:36:23, and for ObsID 10 679 (below), $t = 0$ corresponds to 2009 March 13, 06:47:57. The intermittent nature of EV Lac's flaring behaviour is evident.

We discuss the application of our model to this data set and the relevant results in Section 6.

3 HIDDEN MARKOV MODELS

We begin with a brief review of discrete-time HMMs, in order to present the relevant theory and notation required to understand the models and methods developed in this paper. A readable, but more comprehensive, introduction to HMMs can be found in Zucchini

[2] The data sets, obtained on 2001 September 19 (ObsID 01885; 100.02 ks) and 2009 Mar 13 (ObsID 10679; 95.56 ks) are available via the CDC 235 at https://doi.org/10.25574/cdc.235.

[3] We have also explored the sensitivity of our analysis to the choice of passband splitting energy value. We carried out the analysis using other astrophysically meaningful splits such 0.9 keV – which separates a thermal spectrum from being dominated by low- and high-temperature plasma – and found no qualitative effect on the results.

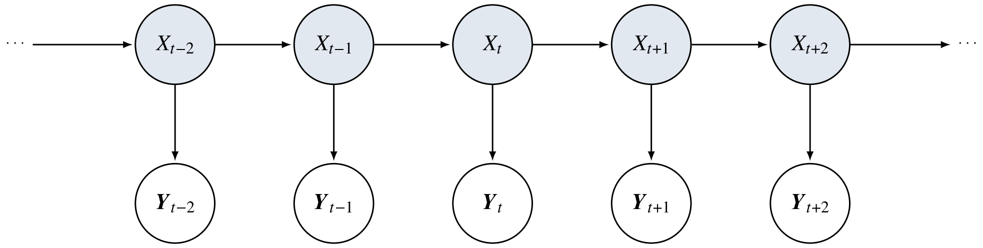

Figure 2. A graphical model representing the standard discretized HMM dependence structure. In this graph, open nodes represent observed quantities and shaded nodes represent unobserved quantities. Generally, an arrow from node X to node Y indicates that the random variables X and Y are not independent, and that the joint distribution of (X, Y) is analysed via the factorization $f_{X,Y}(x, y) = f_{Y|X}(y \mid x) \cdot f_X(x)$ rather than $f_{X,Y}(x, y) = f_{X|Y}(x \mid y) \cdot f_Y(y)$. In the unobserved Markov chain $X_1, X_2, \ldots$, each X_t determines the distribution of its successor X_{t+1} (represented by the forward-pointing arrows). In the observed process $\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots$ each X_t determines the distribution of each $\boldsymbol{Y}_t$ (represented by the downward-pointing arrows), such that, conditional on these determinations, the $\boldsymbol{Y}_t$ are independent (represented by the lack of arrows between the $\boldsymbol{Y}_t$).

et al. (2017) while Cappé, Moulines & Ryden (2005) provide a more advanced treatment.

3.1 Discrete-time hidden Markov models

Heuristically, we employ discrete-time HMMs when we believe that there is an unobserved underlying process governing the distribution of an observed time-series of data at each discrete observation time. For example, we might postulate that a stellar corona is in either a quiescent state or active state at any given time, and that the distribution of observed counts differs between these two states. The underlying state (quiescent or active) is unobserved but governs the distribution of the observed photon counts. Mathematically, the underlying process is modelled as a Markov chain: informally, a sequence of random variables, $X_1, X_2, \ldots$, for which the distribution of any X_t depends on the history of the chain only through the value of X_{t-1}. The variables $X_1, X_2, \ldots$, determine the overall state of the process (e.g. whether the stellar corona is in a quiescent or active state); thus we refer to the X_t as state variables (or simply states).

Inferences about the Markov chain, such as the determination of its values at any time (a process known as state decoding) are performed using only the observed data. Domains in which HMMs commonly appear include meteorology (in which the daily occurrence of rainfall is generated by underlying 'wet' and 'dry' states of nature; Zucchini et al. 2017), animal movement ecology (in which an animal's behavioural states are inferred from telemetry data capturing its physical movements; Langrock et al. 2012b), and finance (in which stock returns are influenced by the underlying state of the economy). In astronomy, Stanislavsky et al. (2020) modelled solar X-ray flux as being generated by underlying 'flaring' and 'non-flaring' states of the sun, as discussed in Introduction.

More formally, the basic discrete-time HMM has two key components. The first component is an unobserved Markov chain, $X_{1:T} = (X_1, \ldots, X_T)$, where each X_t takes values in a common state space $\mathcal{X}$ and the chain is subject to the Markov property,

$$\mathbb{P}(X_t \in A \mid X_{t-1} = x_{t-1}, \ldots, X_t = x_1) = \mathbb{P}(X_t \in A \mid X_{t-1} = x_{t-1}), \tag{1}$$

for all $A \subseteq \mathcal{X}$ (for notational convenience, we start by assuming the X_t are univariate). The second component is a sequence of observed data, $\boldsymbol{Y}_{1:T} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2, \ldots, \boldsymbol{Y}_T)$, where each $\boldsymbol{Y}_t$ takes values in a common observation space $\mathcal{Y}$. For EV Lac, we consider soft and hard passband counts within each time bin; thus each $\boldsymbol{Y}_t$ is bivariate (i.e. a two-component vector), $\mathcal{Y} = \mathbb{R}^2$, and we set $\boldsymbol{Y}_t$ in bold throughout the paper. The two components are subject to the following conditional independence rules:

(i) $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_s$ are conditionally independent given the underlying Markov chain $X_{1:T}$, for any $s \neq t$, and

(ii) the distribution of $\boldsymbol{Y}_t$ depends on $X_{1:T}$ only through the state, X_t, at time index t.

It follows that $\boldsymbol{Y}_t$ and $\boldsymbol{Y}_s$ are conditionally independent given (X_t, X_s) for any $t \neq s$. This means that, conditional on the state of the Markov chain at time index t, the observation $\boldsymbol{Y}_t$ is independent of all other observations; see Fig. 2. Note that (ii) implies that the distribution of each $\boldsymbol{Y}_t$ is fully characterized by the underlying state X_t; often, the distributions of the individual observations, $\boldsymbol{Y}_t$, all belong to the same parametric family (such as a normal distribution), and the state X_t manifests itself in the particular parameters of the distribution of $\boldsymbol{Y}_t$ (such as the mean and variance, in the case of state-dependent normal distributions). In most cases, the state space $\mathcal{X}$ is either finite or a continuum; we describe these cases separately. The notation used here and elsewhere in this paper is summarized in Table 1.

3.2 Discrete-space hidden Markov models

When the state space $\mathcal{X}$ is finite, it is commonly represented as $\mathcal{X} = \{1, \ldots, K\}$ for some $K \in \mathbb{N}$, where each value in $\mathcal{X}$ plays the role of a label for an underlying state of nature (e.g. when $K = 2$, 'flaring' and 'quiescent' can simply be represented as '1' and '2', respectively). In this case, the resulting HMM is referred to as a discrete-space HMM. The specification of a (time-homogeneous) discrete-space HMM consists of three ingredients:

(i) an initial distribution on $\mathcal{X}$, represented by a vector $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_K)$ with $\delta_i = \mathbb{P}(X_1 = i)$,

(ii) a set of transition probabilities, $\gamma_{i,j} = \mathbb{P}(X_{t+1} = j \mid X_t = i)$ for any $t \geq 1$, represented by a $K \times K$ transition matrix, $\boldsymbol{\Gamma}$, with element (i, j) given by $\gamma_{i,j}$, and

(iii) a set of state-dependent distributions, each characterized by a density or mass function $h_k(\mathbf{y} \mid \boldsymbol{\lambda}_k)$ determining the conditional distribution of $\boldsymbol{Y}_t \mid X_t = k$ for any t. Here $\boldsymbol{\lambda}_k$ is a state-specific distributional parameter, which may consist of several components.

Let $\boldsymbol{\eta}$ denote the set of HMM model parameters, including the initial distribution, the transition probabilities, and the parameters of the state-dependent distributions, that is, $\boldsymbol{\eta} = (\boldsymbol{\delta}, \boldsymbol{\Gamma}, \boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K)$.

Table 1. Table of notation used throughout the paper.

w	Time bin width for grouping observations into discrete counts; usually $w = 50$ s
t	Index of time bin
$Y_{t,1}$	Observed soft band count at time index t
$Y_{t,2}$	Observed hard band count at time index t
$\boldsymbol{Y}_t$	Observed bivariate vector of counts (soft and hard band) at time index t (i.e. $\boldsymbol{Y}_t = (Y_{t,1}, Y_{t,2})$)
$\boldsymbol{Y}_{s:s'}$	Collection of observed $\boldsymbol{Y}_t$ from $t = s$ to s'
X_t	State of underlying Markov chain at time index t
$X_{s:s'}$	Collection of underlying states from $t = s$ to s'
$\mathcal{X}$	Underlying state-space which each X_t takes values within
$\boldsymbol{\delta}$	Initial distribution for a discrete-space Markov chain, represented as a vector
$\tilde{\boldsymbol{\delta}}$	Discrete approximation to an initial distribution to a continuous-space Markov chain
$\gamma_{i,j}$	Transition probability from state i to state j for a discrete-space Markov chain
$\boldsymbol{\Gamma}$	Transition matrix for a discrete-space Markov chain
$\tilde{\boldsymbol{\Gamma}}$	Discrete approximation to a transition density of a continuous-space Markov chain
$\lambda_{k,1}$	Parameter for kth state-dependent distribution of soft band count
$\lambda_{k,2}$	Parameter for kth state-dependent distribution of hard band count
$\boldsymbol{\lambda}_k$	Parameter vector for kth state-dependent distribution (i.e. $\boldsymbol{\lambda}_k = (\lambda_{k,1}, \lambda_{k,2})$)
$h_k(\cdot \mid \boldsymbol{\lambda}_k)$	State-dependent density or mass function of $\boldsymbol{Y}_t$ (i.e. conditional on $X_t = k$)
$\boldsymbol{\eta}$	Vector of all unknown parameters in a given model
$L(\boldsymbol{\eta} \mid \mathbf{y}_{1:T})$	Likelihood function (as a function of $\boldsymbol{\eta}$) for a given model
$\delta(\cdot)$	Initial distribution for a continuous-space Markov chain, represented as a density function
$\gamma(\cdot, \cdot)$	Transition density for a continuous-space Markov chain
π	Stationary distribution for a given Markov chain
β_h	Mean emission rate for band h when $X_{t,h} = 0$, scaled by $1/w$ (for $h = 1, 2$)
ϕ_h	Autocorrelation parameter for $X_{t,h}$ (for $h = 1, 2$)
$\boldsymbol{\Phi}$	Autocorrelation matrix with (ϕ_1, ϕ_2) along the diagonal and off-diagonal entries equal to 0
$\varepsilon_{t,1}$	Soft band error/innovation at time index t given by $X_{t,1} - \phi_1 X_{t-1,1}$
$\varepsilon_{t,2}$	Hard band error/innovation at time index t given by $X_{t,2} - \phi_1 X_{t-1,2}$
$\boldsymbol{\varepsilon}_t$	Bivariate error/innovation term at time index t (i.e. $\boldsymbol{\varepsilon}_t = (\varepsilon_{t,1}, \varepsilon_{t,2})$)
σ_h^2	Variance of $\varepsilon_{t,h}$ (for $h = 1, 2$)
ρ	Correlation between $\varepsilon_{t,1}$ and $\varepsilon_{t,2}$
$\mathbf{0}$	Vector of zeros of length 2 (i.e. $\mathbf{0} = (0, 0)$)
$\boldsymbol{\Sigma}$	Covariance matrix with (σ_1^2, σ_2^2) along the diagonal and off-diagonal entries equal to $\sigma_1 \sigma_2 \rho$
$\mathcal{N}(\mu, \sigma^2)$	Univariate normal distribution with mean μ and variance σ^2
$\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$	Bivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$
$\hat{Y}_{t,h}$	Predicted mean (Poisson rate) of distribution of $Y_{t,h}$ (for $h = 1, 2$)
$\hat{X}_{t,h}$	Prediction of $X_{t,h}$ conditional on $\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}$ (for $h = 1, 2$)
$\mathbb{R}$	Set of real numbers
$\mathbb{N}_{\geq 1}$	Set of positive integers
$\mathbb{P}_{\boldsymbol{\eta}}(A)$	Probability of an event A given distributional parameter values $\boldsymbol{\eta}$
$\mathbb{E}_{\boldsymbol{\eta}}[X]$	Expectation of a random variable X given distributional parameter values $\boldsymbol{\eta}$
A_i	Subrectangle i used to partition continuous state space in discrete HMM approximation
$\mathbf{c}_i^*$	Representative point within A_i used to define states in discrete HMM approximation
$\mathbf{1}$	Column vector of ones (i.e. $\mathbf{1} = (1, 1, \ldots, 1)^\top$)
α	Mixing parameter for first component of a two-component finite mixture model
α_j	Mixing parameter for jth component of a K-component finite mixture model
$\boldsymbol{\pi}$	Vector of parameters $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_K)$ in density used for semisupervised classification

The likelihood function for the discrete-space HMM is given by

$$L(\boldsymbol{\eta} \mid \mathbf{y}_{1:T}) = \sum_{x_1=1}^{K} \cdots \sum_{x_T=1}^{K} \left(\delta_{x_1} \cdot h_{x_1}(\boldsymbol{y}_1 \mid \boldsymbol{\lambda}_{x_1}) \prod_{t=2}^{T} \left(\gamma_{x_{t-1},x_t} \cdot h_{x_t}(\boldsymbol{y}_t \mid \boldsymbol{\lambda}_{x_t}) \right) \right). \quad (2)$$

The sums in equation (2) 'marginalize' the unknown state sequence $X_{1:T}$ out of the likelihood by summing over all possible state sequences which could have generated the observed data.

Standard algorithms are available for computing the maximum-likelihood estimate of $\boldsymbol{\eta}$ under equation (2). While the number of terms summed in equation (2) is exponential in T, an efficient algorithm known as the forward algorithm allows the likelihood to be computed in polynomial time; see Appendix A1 for details. Embedding this algorithm within the E-step of the well-known EM algorithm (see Appendix C) produces the Baum–Welch algorithm, which allows for fast maximization of equation (2); see Zucchini et al. (2017) for details. Once the model parameters have been estimated, the forward–backward algorithm (detailed in Appendix A2) can be used to compute posterior state membership probabilities of the form $\hat{p}_{t,k} = \mathbb{P}(X_t = k \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T})$ for each $t = 1, \ldots, T$, and from these, the posterior state membership classifications given by $\mathrm{argmax}_k\, \hat{p}_{t,k}$.

3.3 Continuous-space hidden Markov models

When the state space $\mathcal{X}$ is a continuum (such as $\mathbb{R}$ or, more generally, $\mathbb{R}^d$ for some $d \geq 1$), the resulting HMM is called a continuous-space HMM. In this case, the first two ingredients in the discrete-space

HMM specification are replaced by continuous analogues, while the third is essentially unchanged:

(i) an initial distribution on $\mathcal{X}$, represented by a probability density function $\delta(x)$ satisfying $\mathbb{P}(\boldsymbol{X}_1 \in A) = \int_A \delta(x)\,\mathrm{d}x$ for $A \subseteq \mathcal{X}$,

(ii) a transition density function, $\gamma : \mathcal{X}^2 \rightarrow [0, \infty)$ satisfying $\mathbb{P}(\boldsymbol{X}_{t+1} \in A \mid \boldsymbol{X}_t = \boldsymbol{x}) = \int_A \gamma(\boldsymbol{x}, \boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}'$ for any $t \geq 1$ and $\boldsymbol{x}' \in \mathcal{X}$, and

(iii) a set of state-dependent distributions, each characterized by a density or mass function $h_{\mathbf{x}}(\mathbf{y} \mid \boldsymbol{\lambda}_{\mathbf{x}})$ determining the conditional distribution of $\boldsymbol{Y}_t \mid \boldsymbol{X}_t = \boldsymbol{x}$ for any t. Here, $\boldsymbol{\lambda}_{\mathbf{x}}$ is the parameter specifying the distribution of $\boldsymbol{Y}_t$ given that $\boldsymbol{X}_t = \boldsymbol{x}$; this parameter may consist of several components.

The likelihood function for the continuous-space HMM is

$$L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}) = \int_{\mathcal{X}} \cdots \int_{\mathcal{X}} \delta(\boldsymbol{x}_1) \cdot h_{\boldsymbol{x}_1}(\boldsymbol{y}_1 \mid \boldsymbol{\lambda}_{\boldsymbol{x}_1}) \prod_{t=2}^{T} \gamma(\boldsymbol{x}_{t-1}, \mathbf{x}_t) \cdot h_{\boldsymbol{x}_t}(\boldsymbol{y}_t \mid \boldsymbol{\lambda}_{\boldsymbol{x}_t})\,\mathrm{d}\boldsymbol{x}_{T:1}, \tag{3}$$

where the iterated integrals over $\mathcal{X}$ have replaced the sums in equation (2) and $\mathrm{d}\boldsymbol{x}_{T:1} = \mathrm{d}\boldsymbol{x}_T \cdots \mathrm{d}\boldsymbol{x}_1$.

In both discrete- and continuous-space Markov chains, the corresponding transition probabilities or transition density may induce a stationary distribution for the underlying Markov chain – a distribution π where $X_t \sim \pi$ implies that $X_{t+1} \sim \pi$ (i.e. if one iterate of the chain is marginally distributed according the stationary distribution, all subsequent iterates are also marginally distributed according to π). Under broadly realistic assumptions, the stationary distribution is equal to the asymptotic distribution of the chain, that is, the limiting distribution of X_t as $t \rightarrow \infty$ (e.g. Resnick 2013).

3.4 Approximation to the continuous-space HMM likelihood

In contrast to the situation for the discrete-space HMM, computing the maximum-likelihood estimate under a continuous-space HMM by maximizing equation (3) poses considerable challenges. With the sums over $\{1, \ldots, K\}$ replaced by integrals over $\mathcal{X}$, no efficient algorithms are known that can compute equation (3), let alone maximize it. Fortunately, however, we can approximate equation (3) to arbitrary high level of accuracy by replacing the continuous-space Markov chain with a suitably chosen discrete-space one; this idea originates from the work of Kitagawa (1987) and was developed for state-space models by Langrock (2011). We provide a brief outline of the method and its derivation here, with additional details in Appendix B; see also Langrock (2011) for a complete exposition in the univariate case and Langrock, MacDonald & Zucchini (2012a) for several illustrative examples.

First, we must identify an essential domain A, which is a bounded subset of $\mathcal{X}$ such that $\mathbb{P}(\boldsymbol{X}_t \in A)$ is nearly one for all t (Kitagawa 1987). Next, A must be partitioned into subsets $A_1, \ldots, A_m$ and a representative point, $\boldsymbol{c}_i^*$, chosen for each A_i, for example, $\boldsymbol{c}_i^*$ can be set to the centre of A_i. If all of the A_i are small, then

$$L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}) \approx \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} \left(\mathbb{P}(\boldsymbol{X}_1 \in A_{i_1}) \cdot h_{\boldsymbol{c}_{i_1}^*}\left(\boldsymbol{y}_1 \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_1}^*}\right) \cdot \prod_{t=2}^{T} \left(\mathbb{P}\left(\boldsymbol{X}_t \in A_{i_t} \mid \boldsymbol{X}_{t-1} = \boldsymbol{c}_{i_{t-1}}^*\right) \cdot h_{\boldsymbol{c}_{i_t}^*}\left(\boldsymbol{y}_t \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_t}^*}\right)\right)\right), \tag{4}$$

where the approximation becomes exact as A approaches $\mathcal{X}$ and each of the A_i decrease in size (see Appendix B for details.) Defining the vector $\tilde{\boldsymbol{\delta}} \in \mathbb{R}^m$ and matrix $\tilde{\boldsymbol{\Gamma}} \in \mathbb{R}^{m \times m}$ by the entries

$$\tilde{\delta}_j = \mathbb{P}(\boldsymbol{X}_1 \in A_j) \quad \text{and} \quad \tilde{\gamma}_{i,j} = \mathbb{P}(\boldsymbol{X}_t \in A_j \mid \boldsymbol{X}_{t-1} = \boldsymbol{c}_i^*), \tag{5}$$

the approximation (4) can be written

$$L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}) \approx \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} \left(\tilde{\delta}_{i_1} \cdot h_{\boldsymbol{c}_{i_1}^*}\left(\boldsymbol{y}_1 \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_1}^*}\right) \prod_{t=2}^{T} \left(\tilde{\gamma}_{i_{t-1}, i_t} \cdot h_{\boldsymbol{c}_{i_t}^*}\left(\boldsymbol{y}_t \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_t}^*}\right)\right)\right), \tag{6}$$

where $\boldsymbol{\eta}$ is a vector consisting of the unknown parameters in the state-space model, including the state-dependent parameters $\boldsymbol{\lambda}_{\mathbf{c}_{i,1}^*}, \ldots, \boldsymbol{\lambda}_{\mathbf{c}_{i,T}^*}$ and any parameters associated with the distribution of the underlying Markov chain $\boldsymbol{X}_{1:T}$. If we replace the initial density δ and transition density γ with the discretized functions in equation (5), the approximation in equation (6) is precisely of the form of the discrete-space HMM likelihood given in equation (2), and so, up to the renormalization of $\tilde{\boldsymbol{\delta}}$ and the rows of $\tilde{\boldsymbol{\Gamma}}$, equation (6) is the likelihood of an m-state discrete-space HMM in which the chain being in 'state' i at time index t corresponds to the event that $\boldsymbol{X}_t \in A_i$.

With all elements in the approximation specified in this way, evaluation of equation (6) can proceed using the forward algorithm discussed in Section 3.2. When $\mathcal{X} = \mathbb{R}^d$ for $d > 1$ and the size of the partition m is large, mapping the unordered partition of A to an ordered set of states $\{1, \ldots, m\}$ poses its own challenges. When $d = 2$, this mapping can be accomplished by a pairing function – that is, a bijection from $\mathbb{N}_{\geq 1} \times \mathbb{N}_{\geq 1}$ to $\mathbb{N}_{\geq 1}$. We slightly modify Szudzik's 'Elegant' bijection between $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$ (Szudzik 2006) so that the original function and its inverse have the required domain and range. The modification and its inverse are respectively given by

$$\mathrm{pair}(i, j) = \begin{cases} j^2 - 2j + i + 1, & i \neq \max\{i, j\} \\ i^2 + j - i, & i = \max\{i, j\} \end{cases}$$

and

$$\mathrm{unpair}(j) = \begin{cases} \left(j - g(j)^2, g(j) + 1\right), & j - g(j)^2 - 1 < g(j) \\ \left(g(j) + 1, j - g(j)^2 - g(j)\right), & j - g(j)^2 - 1 \geq g(j) \end{cases},$$

where $g(j) = \lfloor \sqrt{j - 1} \rfloor$.

In practice, one can manually verify that the range of the chosen essential domain is sufficient for the data at hand by inspecting a histogram of the predicted states produced by any state decoding algorithm (see Appendix A2) after the model has been fit (Zucchini et al. 2017).

4 STAGE 1: HMMS FOR FLARING SOURCES

In this section, we propose three new HMMs which are well suited to model flares in stellar coronae. These models are more generally applicable, but because we focus on data sets of flaring stellar light curves (see Section 2), and because other model choices are possible, we caution that it is necessary to consider carefully the particular scenario before adopting these models without suitable modifications. Indeed, we are actively engaged in applying the models to flaring sources other than stars and exploring what generalizations to the models might be appropriate for these application; see Section 7.2 for discussion. All of the models consider photon counts recorded in a sequence of time intervals indexed by t and tabulated into soft passband counts, $Y_{t,1}$, and hard passband counts, $Y_{t,2}$, for $t = 1, \ldots, T$. We start by considering the relative merits of discrete and continuous state spaces as the basis for modelling the flaring behaviour of stars.

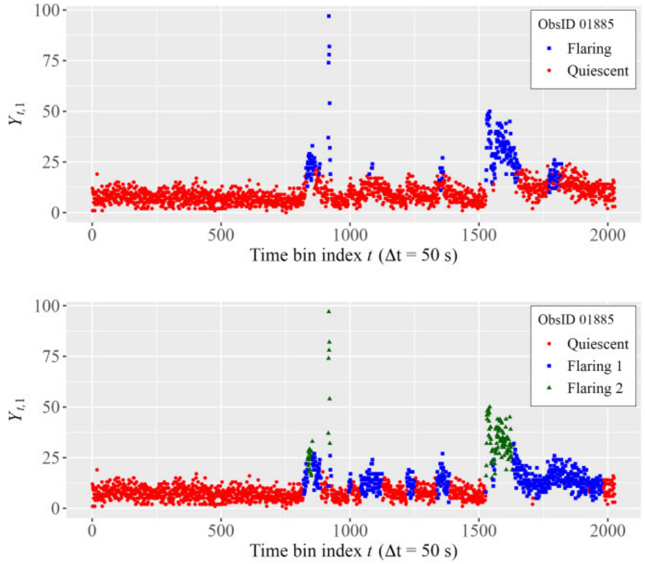

Figure 3. Soft band ObsID 01885 light curve coloured with classifications based on two-state (above) and three-state (below) HMMs fit directly to the observed data $\boldsymbol{Y}_{1:T}$.

4.1 Discrete-space HMMs for flaring stellar coronae

With a discrete state space, a state-dependent bivariate Poisson distribution can be written

$$\boldsymbol{Y}_t \mid X_t = k \sim \text{Poisson}(\lambda_{k,1}) \cdot \text{Poisson}(\lambda_{k,2}), \tag{7}$$

for $t = 1, \ldots, T$ and $k = 1, \ldots, K$, where here and below the notation $\boldsymbol{Y}_t \mid X_t = k \sim \text{Poisson}(\lambda_{k,1}) \cdot \text{Poisson}(\lambda_{k,2})$ indicates that the Poisson distributions of the passbands $Y_{t,1}$ and $Y_{t,2}$, conditional on the event $X_t = k$, are independent for all t. There are many possible alternatives to equation (7) for count data, including combinations of various bivariate Poisson and negative binomial distributions (see Johnson, Kotz & Balakrishnan 1997, for examples) and state-dependent copulas (see Zimmerman, Craiu & Leos-Barajas 2023), all of which induce dependence structures between $Y_{t,1}$ and $Y_{t,2}$. In principle, a two-state HMM could be used to model a star's states as 'quiescent' and 'flaring', roughly in the manner of Stanislavsky et al. (2020). Alternatively, a three-state HMM might split the 'flaring' state into states of rising and falling flaring activity (Esquivel et al. 2024).

We fit the model specified in equation (7) to ObsID 01885 light curve for both $K = 2$ and 3 via maximum likelihood as described in Section 3.2. Fig. 3 illustrates the fitted predicted classifications for each time interval, computed as $\text{argmax}_k\, \hat{p}_{t,k}$, again as described in Section 3.2. Inspection of Fig. 3 (or indeed of Fig. 1) reveals a theoretical defect of using a discrete-space HMM to model the stellar flare process of EV Lac. Under the conditional independence rules of Section 3.1, all observations generated by the same state are independent and identically distributed. Indeed, this implies that the red observations in Fig. 3 must be independent and identically distributed, as are the green and blue ones. This implication is contradicted by the clear temporal trend of the red observations, as well as the sharp rise and fall of the blue ones. Thus, the conditional independence rule is not satisfied and the standard discrete-space HMM is not directly suitable for our data.

This time-series is comprised of jumps between two clearly distinguished levels, pushed by a gradual trend over time (see figs 1 and 2 of Stanislavsky et al. 2020).

4.2 Continuous-space HMMs for flaring stellar coronae

There is no reason to assume that the underlying physical process generating stellar flare activity is binary and is either 'on' or 'off'. Here, we consider a more realistic model that allows the expected photon count at time index t to depend on a continuous underlying process. This enables us to model gradual and/or smooth transitions between a quiescent and an active corona (e.g. with long periods of quiescence interrupted by more intense signals at random intervals). We also weaken the assumption that a single underlying univariate process X_t drives both the hard and soft band photon counts. Specifically, we replace X_t with a bivariate vector $\boldsymbol{X}_t$ whose components $X_{t,1}$ and $X_{t,2}$ may be correlated with each other. We maintain the Markov assumption expressed as a bivariate version of (1).

We specify a Poisson state-space model[4] that satisfies these requirements. First, the state-dependent distribution models flux measurement (i.e. the counts in two passbands) via a Poisson (error) distribution conditional on the underlying $\mathbb{R}^2$-valued state-space variable $\mathbf{X}_t$:

$$\boldsymbol{Y}_t \mid \boldsymbol{X}_t \sim \text{Poisson}\left(w \cdot \beta_1 \cdot e^{X_{t,1}}\right) \cdot \text{Poisson}\left(w \cdot \beta_2 \cdot e^{X_{t,2}}\right). \tag{8}$$

Second, the astrophysical source variability or signal is modelled via an autoregressive process for $\boldsymbol{X}_t$, specified as

$$\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \tag{9a}$$

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix}, \tag{9b}$$

$$\boldsymbol{\varepsilon}_t \overset{\text{iid}}{\sim} \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}), \text{ and} \tag{9c}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}. \tag{9d}$$

The $\boldsymbol{\varepsilon}_t$ term in equation (9a) does not represent observational noise; rather, it represents the random innovation in the underlying source or signal variability. Observational uncertainty, on the other hand, is captured implicitly by the Poisson distribution in equation (8), and not by any explicit additive term in the model. Fitting this Poisson state-space model allows us to go beyond simply fitting the raw light curves. Ultimately, this will allow us to identify time intervals with different statistical behaviours (e.g. quiescence and flaring); see Section 5. Note that the notation $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ in equation (9c) represents a bivariate Gaussian distribution with mean vector equal to $\mathbf{0}$ and covariance matrix equal to $\boldsymbol{\Sigma}$.

The parameters to be estimated in equations (8)–(9d) are $\beta_1, \beta_2 > 0$, the coefficient matrix $\boldsymbol{\Phi}$ with diagonal entries $\phi_1, \phi_2 \in (-1, 1)$, and the covariance matrix $\boldsymbol{\Sigma}$ built up of components $\sigma_1, \sigma_2 > 0$ and $\rho \in (-1, 1)$. The remaining term, w, is the time bin used to group the original photon event list into discrete counts; including w facilitates the study of dependence on bin size (see Appendix D) and also helps to avoid numerical underflow in the estimation process.

Under the model in equations (8)–(9d), the expected photon counts $\mathbb{E}\left[Y_{t,1}\right]$ and $\mathbb{E}\left[Y_{t,2}\right]$ at time index t in the soft and hard bands are monotone increasing functions of $X_{t,1}$ and $X_{t,2}$, respectively. The parameter β_h is proportional to the expected Poisson photon count when $X_{t,h} = 0$. (Since $X_{t,h}$ can take on negative values, $X_{t,h} = 0$

[4] The term 'state-space model' – unlike 'HMM' – is not consistently defined in the literature. Here, we simply regard state-space models as those with observation processes (partially) driven by some hidden linear state process defined on a continuous state space. In other domains such as control theory, this term commonly refers to more specific models in which the observation process is itself a linear function of the state process.

does not necessarily correspond to a state of particularly low or high flaring activity.) The coefficient matrix $\boldsymbol{\Phi}$ determines the extent to which each $X_{t,h}$ is correlated with its predecessor, $X_{t-1,h}$. A slight generalization of equation (9b) allows the off-diagonal entries of $\boldsymbol{\Phi}$ to be nonzero, thereby allowing $X_{t,1}$ to depend on $X_{t-1,2}$ and vice versa (see Section 7).

The state process $\boldsymbol{X}_t$ of the model described by equations (9a)–(9d) is a first-order vector autoregressive process, denoted as a VAR(1) process in the statistical literature. VAR models are commonly applied in areas such as mathematical finance, where they play important roles in stochastic volatility modelling (e.g. Primiceri 2005).

To compute the (approximate) maximum-likelihood estimate under the model in equations (8)–(9d), we maximize the discrete-state-space approximation to the likelihood; see equation (6). Because the state space is $\mathbb{R}^2$, it is convenient – although not strictly necessary – to choose the essential domain A to be a rectangle. Similarly, we partition A into a large number of subrectangles, $A_1, \ldots, A_m$, and set the representative point, $\boldsymbol{c}_i^*$, of each to be its centre.

To numerically optimize equation (6), we use a parallelized version of the popular limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (L-BFGS) routine as implemented in the OPTIMPARALLEL package (Gerber & Furrer 2019) within R. We prefer to use unconstrained optimization to avoid numerical issues caused by parameter inputs lying on the boundaries of their respective domains; thus instead of optimizing the parameters ϕ_1, ϕ_2 and ρ in the approximate likelihood over $(-1, 1)$, we optimize $\tanh^{-1}(\phi_1)$, $\tanh^{-1}(\phi_2)$, and $\tanh^{-1}(\rho)$ over $\mathbb{R}$, and then transform the optimizing values back to their natural domain via the inverse function $x \mapsto \tanh(x)$. Similarly, we optimize $\log \beta_1$, $\log \beta_2$, $\log \sigma_1$, and $\log \sigma_2$ over $\mathbb{R}$, and replace the results with their exponentiated values.

The (approximate) maximum-likelihood estimates may be slightly biased due to small sample sizes. (Maximum-likelihood estimates are asymptotically unbiased for most 'smooth' models, but are generally not unbiased with finite samples.) Similarly, with a small sample size the negative Hessian matrix of the log-likelihood function evaluated at the maximum-likelihood estimate may yield an inadequate approximation to the Fisher information matrix, which is normally used to produce confidence intervals. In order to remedy both issues, we appeal to the parametric bootstrap, which allows us to estimate simultaneously the standard errors of parameter estimates and their biases (Efron & Tibshirani 1993).

Specifically, after computing the maximum-likelihood estimate of the parameters, $\hat{\boldsymbol{\eta}}_{\text{mle}}$, using the actual data, $\boldsymbol{Y}_{1:T}$, we independently generate B replicate data sets, $\boldsymbol{Y}_{1:T}^{(1)}, \ldots, \boldsymbol{Y}_{1:T}^{(B)}$, under the model, each with parameter fixed at $\hat{\boldsymbol{\eta}}_{\text{mle}}$. In the context of an HMM, this requires first simulating the underlying state sequences, $\boldsymbol{X}_{1:T}^{(1)}, \ldots, \boldsymbol{X}_{1:T}^{(B)}$, and then generating each $\boldsymbol{Y}_t^{(b)} \mid \boldsymbol{X}_t^{(b)}$ according to the conditional distribution (8). For each $b = 1, \ldots, B$, we then refit the model using $\boldsymbol{Y}_{1:T}^{(b)}$ to produce a replicate estimate, $\hat{\boldsymbol{\eta}}_{\text{bs}}^{(b)}$. Next, we estimate the bias $\boldsymbol{b}$ and covariance matrix $\mathbf{C}$ of the maximum-likelihood estimator via

$$\hat{\boldsymbol{b}}_{\text{bs}} = \bar{\hat{\boldsymbol{\eta}}}_{\text{bs}} - \hat{\boldsymbol{\eta}}_{\text{mle}} \tag{10}$$

and

$$\hat{\mathbf{C}}_{\text{bs}} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\hat{\boldsymbol{\eta}}_{\text{bs}}^{(b)} - \bar{\hat{\boldsymbol{\eta}}}_{\text{bs}} \right) \left(\hat{\boldsymbol{\eta}}_{\text{bs}}^{(b)} - \bar{\hat{\boldsymbol{\eta}}}_{\text{bs}} \right)^{\top}, \tag{11}$$

where

$$\bar{\hat{\boldsymbol{\eta}}}_{\text{bs}} = \frac{1}{B} \sum_{b=1}^{B} \hat{\boldsymbol{\eta}}_{\text{bs}}^{(b)} \tag{12}$$

is the mean of the bootstrap replicate estimates. Finally, the bootstrap-corrected estimate is $\hat{\boldsymbol{\eta}}_{\text{corr}} = \hat{\boldsymbol{\eta}}_{\text{mle}} - \hat{\boldsymbol{b}}_{\text{bs}}$, and standard errors for the components of $\hat{\boldsymbol{\eta}}_{\text{corr}}$ are equal to the square roots of the diagonal elements of $\hat{\mathbf{C}}_{\text{bs}}$. Approximate 95 per cent confidence intervals for the components are computed with these standard errors.

We conducted a brief simulation study to confirm the veracity of the bootstrap estimates and errors. We simulated data under Model 2 (see Section 4.3.2 below) with a pre-specified $\boldsymbol{\eta} = (\phi_1, \sigma_1, \sigma_2, \beta_1, \beta_2)$ chosen relatively close to the values given in the second column of Table 3. Choosing $B = 100$, we independently repeated the bootstrapping procedure 100 times, producing 100 bootstrap 95 per cent confidence intervals centred around 100 bias-corrected maximum-likelihood estimates. The coverage probabilities for the five parameters (i.e. the number of times each true parameter $\phi_1, \ldots, \beta_2$ fell inside the bootstrap confidence intervals, divided by 100) were 0.93, 0.97, 0.92, 0.92, and 0.91, respectively, which all agree with the expected value of 0.95 at the 95 per cent confidence level.

4.3 Three state-space models for flaring stellar coronae

While the Poisson state-space model in equations (8)–(9d) includes features well suited to stellar flare data, it may be more general than necessary; for example, it is not immediately clear that separate underlying processes, $X_{t,1}$ and $X_{t,2}$, are necessary for the hard and soft bands. We therefore consider two special cases of the model, the first itself a special case of the second, before considering the model in equations (8)–(9d) in its full generality as a third model. Thus, the three models we consider form a nested sequence. For each model, we first provide a stochastic representation, and then give the initial distribution (as characterized by $\tilde{\delta}_j$, for $j \in \{1, \ldots, m\}$) and transition probabilities (as characterized by $\tilde{\gamma}_{i,j}$, for $i, j \in \{1, \ldots, m\}$) of the associated discrete-space HMM approximation to the continuous-space model. This involves expressing both the $\tilde{\delta}_j$ and the $\tilde{\gamma}_{i,j}$ as functions of the parameters involved with the stochastic representation of the underlying state process.

Note that the initial density plays a relatively minor role in the likelihood, and that its impact diminishing as T grows. We follow Langrock (2011) and use the stationary distribution of the state process, $\boldsymbol{X}_{1:T}$, for the initial distribution $\tilde{\delta}_j$. Statistically, this is tantamount to assuming that the distribution of the states that the star inhabits is in equilibrium, and is not evolving over time.[5] The transition probabilities $\tilde{\gamma}_{i,j}$ are derived from the stochastic representation of the model.

4.3.1 Model 1: AR(1) process

To reduce the underlying state process to one dimension, we set $X_{t,1} = X_{t,2} =: X_t$ for all t, in which case the latent process reduces to a univariate first-order autoregressive process, denoted as an AR(1) process for short. The entire state-space model can be written in the simplified form

$$\begin{aligned} \boldsymbol{Y}_t \mid X_t &\sim \text{Poisson}\left(w \cdot \beta_1 \cdot e^{X_t}\right) \cdot \text{Poisson}\left(w \cdot \beta_2 \cdot e^{X_t}\right), \\ X_t &= \phi X_{t-1} + \varepsilon_t, \text{ and} \\ \varepsilon_t &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma^2\right). \end{aligned} \tag{13}$$

[5] This is a reasonable choice for a steadily flaring star like EV Lac, which has not shown evidence of drastic changes in X-ray luminosity during observations over the past several decades (Huenemoerder et al. 2010). This choice is also supported by the steadiness of the spectra in the quiescent and flaring states that we find *post facto* across epochs (see Section 6.2.3).

The vector of unknown parameters for Model 1, $\boldsymbol{\eta}_{\mathrm{M1}} = (\phi, \sigma, \beta_1, \beta_2)$, is fit to the data.

Under Model 1, $X_t = \phi X_{t-1} + \varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, and it can be shown that this process admits a stationary distribution if and only if $\phi \in (-1, 1)$, whence the stationary distribution is given by the $\mathcal{N}\left(0, \sigma^2/(1-\phi^2)\right)$ distribution. Thus, if $A_j = [a_j, b_j]$, then the initial distribution for the discrete-space HMM approximation of Model 1 is taken to be the vector $\tilde{\boldsymbol{\delta}}$ comprised of entries

$$\tilde{\delta}_j = \mathbb{P}(X_t \in A_j) = G_X(b_j) - G_X(a_j), \tag{14}$$

where

$$G_X(x) = \int_{-\infty}^{x} \sqrt{\frac{1-\phi^2}{2\pi\sigma^2}} \exp\left\{-\frac{t^2\left(1-\phi^2\right)}{2\sigma^2}\right\} \mathrm{d}t \tag{15}$$

is the cumulative distribution function (cdf) of the $\mathcal{N}\left(0, \sigma^2/(1-\phi^2)\right)$ distribution.

The transition density $\gamma(x_{t-1}, \cdot)$ for Model 1 is defined as the conditional density of $X_t \mid (X_{t-1} = x_{t-1})$. Under this model, it can be shown that $X_t \mid (X_{t-1} = x_{t-1}) \sim \mathcal{N}(\phi x_{t-1}, \sigma^2)$, and so, if c_i^* is the representative point chosen within the interval A_i, then the transitions probabilities between states $\gamma_{i,j} = \mathbb{P}(X_t \in A_j \mid X_{t-1} \in A_i)$ are approximated by

$$\tilde{\gamma}_{i,j} = \mathbb{P}(X_t \in A_j \mid X_{t-1} = c_i^*) = F_{X,i}(b_j) - F_{X,i}(a_j), \tag{16}$$

where

$$F_{X,i}(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(t-\phi c_i^*)^2}{2\sigma^2}\right\} \mathrm{d}t \tag{17}$$

is the cdf of the $\mathcal{N}\left(\phi c_i^*, \sigma^2\right)$ distribution. The $\tilde{\gamma}_{i,j}$ are then taken as the entries of the transition matrix in the discrete-space HMM approximation of the model.

4.3.2 Model 2: VAR(1) process on a line

Model 1 can be viewed as a special case of the general Poisson state-space model (8)–(9d), where $X_{t,1}$ is forced to be equal to $X_{t,2}$ with probability 1 for all t. In Model 2, we relax this restriction and allow $X_{t,2}$ to depend positively and linearly on $X_{t,1}$; specifically, we set $X_{t,2} = \sigma_2 X_{t,1}/\sigma_1$ with probability 1, where each $\sigma_h > 0$ is given by $\sigma_h^2 = \mathrm{Var}\left(X_{t,h} \mid X_{t-1,h}\right)$ for all t. (The assumption of stationarity implies that this variance does not depend on t.) Formally, this can be written as a bivariate state-space model where the $\boldsymbol{X}_t$ follow the degenerate distribution implied by

$$\begin{aligned}
\boldsymbol{X}_t &= \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \\
\boldsymbol{\Phi} &= \begin{bmatrix} \phi & 0 \\ 0 & \phi \end{bmatrix}, \text{ and} \\
\boldsymbol{\varepsilon}_t &\overset{\text{iid}}{\sim} \lim_{\rho \to 1} \mathcal{N}_2\left(\mathbf{0}, \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}\right).
\end{aligned} \tag{18}$$

The bivariate distribution for $\boldsymbol{\varepsilon}_t$ lacks a density with respect to Lebesgue measure on $\mathbb{R}^2$, but admits a density on the line $y = \sigma_2 x/\sigma_1$. However, it is more convenient to write the state-space model entirely in terms of the univariate state process $X_t := X_{t,1}$ as

$$\begin{aligned}
\boldsymbol{Y}_t \mid X_t &\sim \text{Poisson}\left(w \cdot \beta_1 \cdot e^{X_t}\right) \cdot \text{Poisson}\left(w \cdot \beta_2 \cdot e^{\sigma_2 X_t/\sigma_1}\right), \\
X_t &= \phi X_{t-1} + \varepsilon_t, \text{ and} \\
\varepsilon_t &\overset{\text{iid}}{\sim} \mathcal{N}\left(0, \sigma_1^2\right).
\end{aligned} \tag{19}$$

The vector of unknown parameters for Model 2, $\boldsymbol{\eta}_{\mathrm{M2}} = (\phi, \sigma_1, \sigma_2, \beta_1, \beta_2)$, is fit to the data.

In the bivariate formulation (18), $X_{t,1}$ lies within the interval $[a, b] \subset \mathbb{R}$ if and only if $X_{t,2}$ lies within $[\sigma_2 a/\sigma_1, \sigma_2 b/\sigma_1]$ with probability 1. Thus, the transition probabilities for $X_{t,2}$ are determined by those of $X_{t,1}$ alone, as is the initial distribution of $X_{t,2}$ (since we assume $X_{t,1}$ – and therefore $X_{t,2}$ – is stationary). It follows that the initial distribution $\tilde{\boldsymbol{\delta}}$ and the transition probabilities $\tilde{\gamma}_{i,j}$ for Model 2 are exactly the same as those in Model 1, but with σ replaced by σ_1; effectively, the only difference between Models 1 and 2 is the inclusion of σ_2/σ_1 in the state-dependent Poisson distribution corresponding to the hard-band photons. For the process $X_t = \phi X_{t-1} + \varepsilon_t$ to be stationary, we again require that $\phi \in (-1, 1)$.

4.3.3 Model 3: uncorrelated VAR(1) process

Model 3 further generalizes Model 2 by removing the restriction that $X_{1,t}$ and $X_{2,t}$ depend on each other linearly. In particular, Model 3 allows $X_{1,t}$ and $X_{2,t}$ to move freely in their own 'directions', but ensures dependence between them by way of correlated errors. Specifically,

$$\begin{aligned}
\boldsymbol{Y}_t \mid \boldsymbol{X}_t &\sim \text{Poisson}\left(w \cdot \beta_1 \cdot e^{X_{t,1}}\right) \cdot \text{Poisson}\left(w \cdot \beta_2 \cdot e^{X_{t,2}}\right), \\
\boldsymbol{X}_t &= \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t, \\
\boldsymbol{\Phi} &= \begin{bmatrix} \phi_1 & 0 \\ 0 & \phi_2 \end{bmatrix}, \\
\boldsymbol{\varepsilon}_t &\overset{\text{iid}}{\sim} \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma}), \text{ and} \\
\boldsymbol{\Sigma} &= \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{bmatrix}.
\end{aligned} \tag{20}$$

The vector of unknown parameters for Model 3, $\boldsymbol{\eta}_{\mathrm{M3}} = (\phi_1, \phi_2, \sigma_1, \sigma_2, \beta_1, \beta_2, \rho)$, is fit to the data.

Model 3 includes two more parameters than Model 2, namely, $\rho \in (-1, 1)$ and $\phi_2 > 0$. In contrast to Model 2, here densities with respect to $\mathbb{R}^2$ exist for the bivariate conditional and stationary distributions of the $\boldsymbol{X}_t$. Since $\boldsymbol{X}_t$ can lie within any open set of $\mathbb{R}^2$ with positive probability, the resulting initial distribution and transition probabilities in the discrete-space HMM approximation to the model must be derived anew.

Under Model 3, the existence of a stationary distribution for the process $\boldsymbol{X}_t = \boldsymbol{\Phi}\boldsymbol{X}_{t-1} + \boldsymbol{\varepsilon}_t$ requires that $\phi_1, \phi_2 \in (-1, 1)$. The corresponding distribution is well known (e.g. Hamilton 2020) and is given by the $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Lambda})$ distribution, where $\mathrm{vec}\,(\boldsymbol{\Lambda}) = (\boldsymbol{I} - \boldsymbol{\Phi} \otimes \boldsymbol{\Phi})^{-1}\mathrm{vec}\,(\boldsymbol{\Sigma})$, $\mathbf{I}$ is the 4×4 identity matrix, $\otimes$ is the Kronecker product between matrices, and $\mathrm{vec}\,(\cdot)$ is the vectorization operator that stacks the columns of an $m \times n$ matrix into a $mn \times 1$ vector. Thus, if $A_j = [a_{j,1}, b_{j,1}] \times [a_{j,2}, b_{j,2}]$ which, rather than an interval in $\mathbb{R}$ as in Models 1 and 2, is now a rectangle in $\mathbb{R}^2$, then the initial distribution for the discrete-space HMM approximation of Model 3 is taken to be the vector $\tilde{\boldsymbol{\delta}}$ comprised of entries

$$\begin{aligned}
\tilde{\delta}_j &= \mathbb{P}(\boldsymbol{X}_t \in A_j) \\
&= G_{\boldsymbol{X}}(a_{j,2}, b_{j,2}) - G_{\boldsymbol{X}}(a_{j,2}, b_{j,1}) - G_{\boldsymbol{X}}(a_{j,1}, b_{j,2}) \\
&\quad + G_{\boldsymbol{X}}(a_{j,1}, b_{j,1}),
\end{aligned} \tag{21}$$

where

$$G_{\boldsymbol{X}}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \frac{1}{2\pi\sqrt{|\det \boldsymbol{\Lambda}|}} \exp\left\{\boldsymbol{t}^\top \boldsymbol{\Lambda}^{-1} \mathbf{t}\right\} \mathrm{d}\boldsymbol{t} \tag{22}$$

is the cdf of the $\mathcal{N}_2(\mathbf{0}, \boldsymbol{\Lambda})$ distribution.

The transition density $\gamma(\boldsymbol{x}_{t-1}, \cdot)$ for Model 3 is now defined as the conditional density of $\boldsymbol{X}_t \mid (\boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1})$. Under this model, it can be shown that $\boldsymbol{X}_t \mid (\boldsymbol{X}_{t-1} = \boldsymbol{x}_{t-1}) \sim \mathcal{N}_2(\boldsymbol{\Phi}\boldsymbol{x}_{t-1}, \boldsymbol{\Sigma})$ and so, if $\boldsymbol{c}_i^*$ is the representative point chosen within the rectangle A_i, then the transitions between states $\gamma_{i,j} = \mathbb{P}(\boldsymbol{X}_t \in A_j \mid \boldsymbol{X}_{t-1} \in A_i)$ are approximated by

$$\begin{aligned}\tilde{\gamma}_{i,j} &= \mathbb{P}(\boldsymbol{X}_t \in A_j \mid \boldsymbol{X}_{t-1} = \boldsymbol{c}_i^*) \\ &= F_{\boldsymbol{X},i}(a_{j,2}, b_{j,2}) - F_{\boldsymbol{X},i}(a_{j,2}, b_{j,1}) - F_{\boldsymbol{X},i}(a_{j,1}, b_{j,2}) \\ &\quad + F_{\mathbf{X},i}(a_{j,1}, b_{j,1}), \end{aligned} \tag{23}$$

where

$$F_{\boldsymbol{X},i}(x_1, x_2) = \\ \int_{-\infty}^{x_1}\int_{-\infty}^{x_2} \frac{1}{2\pi\sqrt{|\det \boldsymbol{\Sigma}|}} \exp\left\{\left(\boldsymbol{t} - \boldsymbol{\Phi c}_i^*\right)^\top \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{t} - \boldsymbol{\Phi c}_i^*\right)\right\} \mathrm{d}\boldsymbol{t} \tag{24}$$

is the $\mathcal{N}_2(\boldsymbol{\Phi}\mathbf{c}_i^*, \boldsymbol{\Sigma})$ cdf. The bivariate normal cdfs (22) and (24) can be computed efficiently using any statistical software package.

4.4 State-space model selection

The three models discussed in Section 4.3 are nested within each other: Model 1 is a special case of Model 2 subject to the constraint $\sigma_1 = \sigma_2$, and Model 2 is a special case of Model 3 subject to the constraints $\phi_1 = \phi_2$ and $\rho = 1$. Thus, any two of these models can, at least in principle, be compared using a likelihood ratio test (LRT). Under certain conditions, if the data are generated under the 'simpler' of the two models being compared (i.e. the model with fewer parameters), the LRT statistic is asymptotically[6] model distributed $\chi^2_{(\nu)}$ with degrees of freedom ν equal to the difference in the number of parameters between the two models. Under certain conditions (e.g. Protassov et al. 2002), this result allows a p-value to be computed; when the LRT statistic is sufficiently large relative to its asymptotic $\chi^2_{(\nu)}$ distribution, a small p-value is obtained and we can conclude that the data are inconsistent with the simpler model. The LRT statistic is equal to −2 times the difference of the maximized log-likelihood functions of the two models under comparison. Thus, we reject the smaller model when the larger model sufficiently improves the fit to a degree as measured by the log-likelihood function.

Among the conditions required for the LRT's asymptotic $\chi^2_{(\nu)}$ distribution are that (i) the models under comparison are nested and (ii) the parameters of the smaller model are not constrained to be on the boundary of the set of possible parameter values under the larger model. These conditions are met for Models 1 and 2 and the standard LRT is thus a suitable means of comparing them. Unfortunately, the comparison of Models 2 and 3 does not satisfy the second of these conditions because one parameter in the smaller Model 2 lies on the boundary of the parameter space of the larger Model 3 (i.e. $\rho = 1$). In fact, the asymptotic distribution of the LRT statistic is not known in this case. While Self & Liang (1987) provided a generalized LRT statistic that helps to account for such situations, its implementation can be computationally difficult.

When the choice between Models 2 and 3 is not clear from the results of the model estimation procedure (as is the case for the EV Lac data; see Section 6.1), one can again use the parametric bootstrap, this time to approximate the finite-sample distribution of the LRT statistic by way of simulations. Assuming that model fitting produces the MLEs $\hat{\boldsymbol{\eta}}_{\mathrm{M2}}$ for Model 2 and $\hat{\boldsymbol{\eta}}_{\mathrm{M3}}$ for Model 3, this bootstrap procedure generates a large number B of independent replicate data sets $\boldsymbol{Y}^{(1)}_{1:T}, \ldots, \boldsymbol{Y}^{(B)}_{1:T}$ under Model 2 with parameter $\hat{\boldsymbol{\eta}}_{\mathrm{M2}}$. For each $b = 1, \ldots, B$, both Models 2 and 3 are fit to $\boldsymbol{Y}^{(b)}_{1:T}$, producing the respective MLEs $\hat{\boldsymbol{\eta}}^{(b)}_{\mathrm{M2}}$ and $\hat{\boldsymbol{\eta}}^{(b)}_{\mathrm{M3}}$. The bootstrapped LRT statistics $\hat{\psi}^{(b)} = -2(\ell_{\mathrm{M2}}(\hat{\boldsymbol{\eta}}^{(b)}_{\mathrm{M2}}) - \ell_{\mathrm{M3}}(\hat{\boldsymbol{\eta}}^{(b)}_{\mathrm{M3}}))$ are computed, where ℓ_{M2} and ℓ_{M3} are the log-likelihood functions for Models 2 and 3, respectively. The statistics $\hat{\psi}^{(1)}, \ldots, \hat{\psi}^{(B)}$ are then used to construct an approximate distribution $\hat{F}_\psi$, perhaps using a kernel density estimate (KDE; see Section 5.1). This distribution is used in place of the $\chi^2_{(\nu)}$ distribution to compute a p-value. Specifically, Model 2 can be rejected in favour of Model 3 at the 95 per cent confidence level if the LRT statistic produced from the original data, $\hat{\psi} = -2(\ell_{\mathrm{M2}}(\hat{\boldsymbol{\eta}}_{\mathrm{M2}}) - \ell_{\mathrm{M3}}(\hat{\boldsymbol{\eta}}_{\mathrm{M3}}))$, is such that $1 - \hat{F}_\psi(\hat{\psi}) < 0.05$.

In addition to overcoming theoretical roadblocks associated with the standard LRT approach, the bootstrap technique helps to account for potential numerical inaccuracies (e.g. stemming from the discrete-space HMM approximation of the state-space likelihood or its optimization, which is especially relevant when the dimension of state space $\mathcal{X}$ is greater than 1). Because the same numerical inaccuracies affect the LRT statistic as computed on the data and as computed on the bootstrap replicates, the bootstrap provides the null distribution of the LRT statistic as it is computed. This allows us to define the statistic to be as computed (including potential numerical inaccuracies) and correctly calibrate its null distribution and the p-value. Specifically, the Monte Carlo nature of the bootstrapped p-values takes the entire approximation procedure into account, whereas the standard LRT approach assumes the use of genuine log-likelihood functions which are perfectly optimized in the involved calculations.

Having fit the state-space model, standard HMM algorithms (see Appendix A) allow one to decode the observations, that is, to make predictions, $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_T$, of the underlying states, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_T$. With a continuous-space HMM, predictions take values in the set of representative points $\{\boldsymbol{c}_1^*, \ldots, \boldsymbol{c}_m^*\}$ defined in the discrete approximation to the continuous-space likelihood, see Section 3.4.

5 STAGE 2: CLASSIFYING LIGHT CURVES INTO FLARING AND QUIESCENT INTERVALS

Our Stage 1 HMMs use continuous underlying processes to model stellar flare activity (see Section 4.2). In practice, however, we also wish to identify those time intervals when the star is in its quiescent state and those when it is in its flaring state. In this section, we introduce our Stage 2 analysis, which uses a finite mixture model to classify the $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_T$ fitted in Stage 1 into the quiescent and flaring states.

We consider two scenarios: semisupervised and unsupervised classification. The semisupervised scenario applies in cases where we are able to identify a subsample of size m of the predicted states, $A_{\mathrm{q}} = \{\hat{\boldsymbol{X}}_{t_1}, \ldots, \hat{\boldsymbol{X}}_{t_m}\}$, where m is reasonably large and the subsample is assumed to arise from a period of quiescence. Identifying a quiescent subsample invariably involves a degree of subjectivity (e.g. through visual inspection). We refer to this scenario as semisupervised because some, but not all, of the data is assumed to be classified a priori. If there is a clearly identifiable interval of quiescence, A_{q} can be selected using a range of time bins where the light curves appear to be in equilibrium and do not exhibit flaring behaviour. In the unsupervised scenario, we do not have such a subsample.

In both the semisupervised and unsupervised scenarios, we propose to model the full set of Stage 1 predicted states, $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_T$, as a mixture of two distributions, one corresponding to the quiescent

[6] The distribution function of the LRT statistic converges pointwise to that of a $\chi^2_{(\nu)}$ random variable as the size of the data set increases (e.g. as the total time duration of the light curve increases). It is in this sense that the LRT statistic is asymptotically $\chi^2_{(\nu)}$-distributed. This assumes that the necessary theoretical conditions are met (e.g. Protassov et al. 2002) and that the data are generated under the simpler

state and the other corresponding to the flaring state.[7] This modelling approach is corroborated by the histogram of the EV Lac state predictions shown in Section 6.2. Formally, we assign the label ‘1’ to the quiescent distribution and ‘2’ to the flaring distribution, and for the purpose of classification, suppose

$$\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_T \overset{\text{iid}}{\sim} \alpha \cdot F_1 + (1-\alpha) \cdot F_2, \tag{25}$$

where F_1 and F_2 are cdfs and $\alpha \in (0, 1)$ is a mixing parameter, all to be inferred from the data. The mixing parameter corresponds to the proportion of time that the star spent in the quiescent state. Model (25) can be equivalently represented by introducing a sequence of latent variables, $Z_1, \ldots, Z_T \overset{\text{iid}}{\sim} \text{Bernoulli}(\alpha)$ and declaring

$$\hat{\boldsymbol{X}}_t \mid Z_t = k \sim F_k, \quad \text{for each } t \text{ and for } k \in \{1, 2\}. \tag{26}$$

Note that neither of these model representations accounts for the autocorrelation (or more generally, the time-series nature) of $\hat{\boldsymbol{X}}_1, \ldots, \hat{\boldsymbol{X}}_T$ implied by the Stage 1 HMMs (e.g. equations 9d, 13, 19, and 20) and observed in the actual EV Lac fits (see e.g. Fig. 5). Instead, we assume that temporal characteristics are captured by the Stage 1 HMM fit, and here we merely aim to classify the light curve into flaring and quiescent intervals.

For simplicity, we assume henceforth that as for Models 1 and 2, the predicted states are univariate, although our theory generalizes to higher dimensional state predictions (as in Model 3). While mixture models often involve component distributions belonging to the same parametric family – normal distributions or other exponential family distributions are especially popular – we consider a less rigid approach to the choices of F_1 and F_2. Ultimately, the estimated probability that the star is in a flaring state depends on the relative size of $f_1(x)$ and $f_2(x)$ at each value of x, where f_1 and f_2 are the probability density functions corresponding to F_1 and F_2, respectively. The choice of f_1 and f_2 is particularly influential for ranges of x at the transition between states, where $f_1(x)$ and $f_2(x)$ are both moderate and are both well above zero; thus, the choice of densities is important, and poor approximations using standard parametric families can potentially yield inaccurate flaring state probabilities for such x.

Note that in both the semisupervised and unsupervised procedures, we are not concerned with overfitting the relevant mixture distributions to the data, as each fitted distribution pertains specifically to the predicted states output by a particular fitted state-space model and are not intended to be used elsewhere.

5.1 Semisupervised classification

There is a distinct advantage in the semisupervised scenario where A_{q} can be used to form a robust non-parametric estimate of $f_1(x)$. Under the mixture model, we have $\hat{X}_{t_j} \sim F_1$ for $j = 1, \ldots, m$ (i.e. for $\hat{X}_{t_j} \in A_{\text{q}}$) and we can use a KDE, $\hat{f}_1(x)$, to approximate $f_1(x)$. The KDE essentially traces out a smoothed version of the histogram of the sample A_{q}.

The flaring component density $f_2(x)$, on the other hand, does not yield as easily to a KDE because an analogous subsample of data known to be from the flaring state is usually unavailable. Instead, we approximate $f_2(x)$ by a step function $\hat{f}_2(x; \boldsymbol{\pi})$ parametrized by the constant value π_k that it takes within a pre-specified bounded interval $[b_{k-1}, b_k)$ for a fixed number K of intervals; that is,

$$\hat{f}_2(x; \boldsymbol{\pi}) = \sum_{k=1}^{K} \frac{\pi_k}{b_k - b_{k-1}} \cdot \mathbb{1}_{x \in [b_{k-1}, b_k)}, \tag{27}$$

where the π_k are unknown non-negative parameters subject to $\sum_{k=1}^{K} \pi_k = 1$. When the intervals $[b_{k-1}, b_k)$ are evenly spaced, $\hat{f}_2$ is essentially a histogram function. We choose $b_K = \sup A$, where A is the essential domain used to approximate the domain of the X_t (see Section 3.4). This is because the values of $\hat{X}_t$ produced by the local decoding algorithm (see Appendix A2) take values in the set of representative points $\{c_1^*, \ldots, c_m^*\} \subseteq A$; thus $\hat{X}_t \in A$ for all t. On the other hand, we assume that the smallest values of X_t are reserved for the quiescent state and thus we choose b_0 as the median of $\hat{f}_1(x)$, although other choices are possible.

The unknown parameters in the model (25) , namely α and $\boldsymbol{\pi}$, can be estimated using the EM algorithm, which is a standard tool for computing maximum-likelihood estimates in finite mixture models (see Dempster, Laird & Rubin 1977) and is easily derived for equation (25) (see Appendix C1). We run the EM algorithm on the subset $A_{\text{r}} = \{\hat{X}_1, \ldots, \hat{X}_T\} \setminus A_{\text{q}}$ of predicted states not used to fit $\hat{f}_1(x)$, so as not to use A_{q} twice in the estimation process; this requires a minor adjustment to the mixing parameter α to account for the proportion of quiescent state data removed (see Appendix C1).

Once the estimation of equation (25) is complete, the estimated posterior probability that each X_t is in a flaring state (i.e. state ‘2’) can be derived using the representation in equation (26), which yields

$$\mathbb{P}(Z_t = 2 \mid \hat{X}_t = \hat{x}_t) = \frac{(1-\hat{\alpha}) \cdot \hat{f}_2(\hat{x}_t; \hat{\boldsymbol{\pi}})}{\hat{\alpha} \cdot \hat{f}_1(\hat{x}_t) + (1-\hat{\alpha}) \cdot \hat{f}_2(\hat{x}_t; \hat{\boldsymbol{\pi}})}, \tag{28}$$

where $\hat{\alpha}$ and $\hat{\boldsymbol{\pi}}$ are the maximum-likelihood estimates computed with the EM algorithm.

5.2 Unsupervised classification

In situations, where there is no subsample of the data that can reasonably be assumed to have arisen from the quiescent state, inference must be fully unsupervised and there is no immediate way to use KDE to approximate $f_1(x)$. In this case, we have found that for the EV Lac data a mixture of three normal distributions provides a reasonable approximation to the distribution of the $\hat{X}_t$: that is,

$$\hat{X}_1, \ldots, \hat{X}_T \overset{\text{iid}}{\sim} \alpha_1 \cdot \mathcal{N}\left(\mu_1, \tau_1^2\right) + \alpha_2 \cdot \mathcal{N}\left(\mu_2, \tau_2^2\right) + \alpha_3 \cdot \mathcal{N}\left(\mu, \tau_3^2\right), \tag{29}$$

where α_1, α_2, and α_3 are non-negative mixing parameters subject to $\sum_{k=1}^{3} \alpha_k = 1$ and each μ_k and τ_k^2 are mean and variance parameters (respectively), all to be estimated. This distribution is also fit using the EM algorithm (see Appendix C2).

In this instance, we assume that one component of the model corresponds to the flaring state, while the remaining two components together correspond to the quiescent state (see Section 6.2.2 for

[7] In the unsupervised scenario, one could, in principle, apply a non-parametric unsupervised clustering method such as k-means (with $k = 2$) to the $\hat{\mathbf{X}}_t$ to classify observations into quiescent and flaring intervals. Such methods have the benefit of being fully automatic, and are easy to implement using built-in routines within any statistical software package. However, quantification of uncertainty for the classifications produced by these ‘black-box’ algorithms are difficult to interpret (and are often not available at all), particularly when there is not a probabilistic model underlying the algorithm. Furthermore, different unsupervised clustering algorithms (e.g. k-means, k-medians, DBSCAN, etc.) use different loss/objective functions and can yield different classifications of the same data; aside from computational complexity, there are few clear reasons for choosing one clustering algorithm over another. Thus, we deploy a more statistical approach, using a likelihood-based finite mixture model.

further discussion in the context of EV Lac). We may assume without loss of generality that $\mu_1 < \mu_2 < \mu_3$, and since a lower X_t corresponds to a lower Poisson intensity for the emission $Y_{t,1}$ (see equation 8), we regard the first two normal distributions in equation (29) as those corresponding to quiescence, with $\alpha_1 + \alpha_2$ representing the proportion of time spent in that state. By using two normal distributions, we are able to better represent the skew in the quiescent distribution. Once equation (29) has been fitted, the posterior probability that each X_t is in a flaring state is given by

$$\mathbb{P}(Z_t \neq 1 \mid \hat{X}_t = \hat{x}_t) = \frac{\hat{\alpha}_3 \cdot f\left(\hat{x}; \hat{\mu}_3, \hat{\tau}_3^2\right)}{\sum_{k=1}^{3} \hat{\alpha}_k \cdot f\left(\hat{x}_t; \hat{\mu}_k, \hat{\tau}_k^2\right)}, \tag{30}$$

where $f(\cdot; \mu, \tau^2)$ is the density of the $\mathcal{N}(\mu, \tau^2)$ distribution.

6 ANALYSIS OF EV LAC

In this section, we illustrate the statistical methods developed in the Sections 3–5 by applying them to the EV Lac data described in Section 2. In particular, we derive a classification of the light curves in Fig. 1 into quiescent and flaring intervals.

6.1 Stage 1: HMM selection and fit for EV Lac

We analysed the two long-duration *Chandra* observations of EV Lac, ObsID 01885 obtained in 2001 and ObsID 10 679 obtained in 2009. For both observations, we used the dispersed data from the combined HEG and MEG arms, and from the combined positive and negative orders, which avoids pileup effects seen during the stronger flares in the zeroth order. We split the data into soft (0.3–1.5 keV) and hard (1.5–8 keV) passbands, and binned them into time bins of $w = 50$ s (see Fig. 1). We also tested the sensitivity of our model fits to these binning schemes by replicating the results using other passbands (i.e. 0.3–0.9 and 0.9–8 keV) and changing the binning phase by 25 s, and found no qualitative differences; see Appendix D for details.

We fit the three state-space models described in Section 4.3 to both observations. For brevity, we present only the fitted models for ObsID 01885; classifications into flaring and quiescent intervals are presented for both observations in Section 6.2. We employed visual diagnostics to determine the parameters of the discretizations of the continuous state spaces. For Model 2, for example, we chose the essential domain $A = [-1.25, 2.65]$ and partitioned A into $m = 40$ evenly spaced subintervals and chose the representative points $\{c_1^*, \ldots, c_{40}^*\}$ as the mid-points of these subintervals; a histogram of the states $\hat{X}_t$ predicted by the model via local decoding shows that this choice of essential domain was conservative in that it easily covers the range of the $\hat{X}_t$ (see the upper panel of Fig. 4). The estimates can be sensitive to the choice of m when m is small and we chose $m = 40$ because this is the approximate number of subintervals at which the parameter estimates and maximized log-likelihood stabilized. Similarly, for Model 3 we chose the essential domain $A = [-1.25, 2.56] \times [-1.75, 3.6]$ (see the lower panel of Fig. 4) and $m = 40^2$.

Bias-corrected parameter estimates and confidence intervals computed using the parametric bootstrap (see Section 4.2) under Models 1, 2, and 3 appear in Tables 2, 3, and 4, respectively. The estimates of the parameters common to the models are broadly consistent with each other, as are their standard errors. The estimates are also very similar to those produced with a passband split at 0.9 keV (omitted for brevity), demonstrating robustness to that choice.

As a byproduct of the optimization procedure used to fit the models, we extracted the values of the maximized log-likelihood

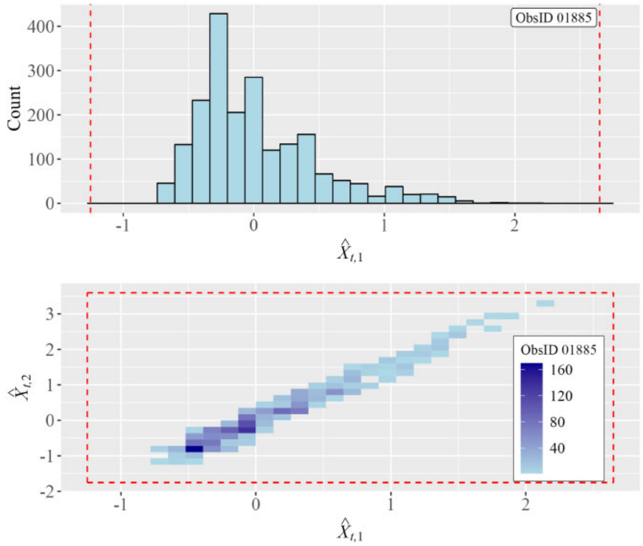

Figure 4. Univariate (top panel) and bivariate (bottom panel) histograms of predicted states $\hat{X}_t$ based on an initial fits of Models 2 and 3 to ObsID 01885; above, the dashed lines enclose the essential domain $A = [-1.25, 2.65]$ chosen for the discrete-space approximation of Model 2, and below, they enclose the essential domain $A = [-1.25, 2.56] \times [-1.75, 3.6]$ chosen for Model 3.

Table 2. Bias-corrected parameter estimates for Model 1 fit to ObsID 01885, with bias correction and standard errors obtained via the parametric bootstrap.

Parameter	Estimate	Standard error	CI (lower)	CI (upper)
ϕ_1	0.987235	0.004579	0.978260	0.996209
σ_1	0.128329	0.006212	0.116155	0.140504
β_1	0.184642	0.045141	0.096165	0.273119
β_2	0.075158	0.018178	0.039529	0.110788

Table 3. Bias-corrected parameter estimates for Model 2 fit to ObsID 01885, with bias correction and standard errors obtained via the parametric bootstrap.

Parameter	Estimate	Standard error	CI (lower)	CI (upper)
ϕ_1	0.979644	0.006456	0.966991	0.992297
σ_1	0.100712	0.004811	0.091282	0.110142
σ_2	0.161689	0.007409	0.147168	0.176210
β_1	0.193817	0.022021	0.150656	0.236978
β_2	0.062417	0.010696	0.041453	0.083380

Table 4. Bias-corrected parameter estimates for Model 3 fit to ObsID 01885, with bias correction and standard errors obtained via the parametric bootstrap.

Parameter	Estimate	Standard error	CI (lower)	CI (upper)
ϕ_1	0.981721	0.008663	0.964742	0.998700
ϕ_2	0.976232	0.007997	0.960558	0.991906
σ_1	0.096086	0.006253	0.083829	0.108342
σ_2	0.171667	0.008918	0.154188	0.189147
β_1	0.206301	0.021047	0.165048	0.247554
β_2	0.066548	0.009203	0.048510	0.084585
ρ	1.000000	0.000000	1.000000	1.000000

Table 5. Maximized log-likelihoods for all three models based on ObsID 01885.

Model	Maximized log-likelihood
Model 1: AR(1) process	−9914.53
Model 2: VAR(1) process on a line	−9455.21
Model 3: Uncorrelated VAR(1) process	−9424.47

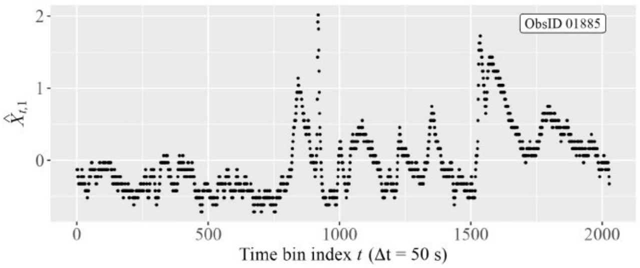

Figure 5. Predicted soft band states $\hat{X}_1, \dots, \hat{X}_{2027}$ for ObsID 01885.

function (6) for each model (shown in Table 5). The standard LRT decisively rejected Model 1 in favour of Model 2, with a test statistic of 918.64 far exceeding the asymptotic $\chi^2_{(1)}$ distribution at the 95 per cent significance level. For a comparison between Models 2 and 3, we turned to the bias-corrected parameter estimates and their corresponding bootstrap standard errors shown in Tables 3 and 4, respectively. These tables show that the correlation parameter ρ in Model 3 is estimated at 1 – precisely its value fixed by Model 2 – with virtually no uncertainty in the estimate (all values have been rounded to six significant figures). Moreover, the remaining parameters shared by Models 2 and 3 are estimated very consistently between the two models, as are their standard errors, and the Model 3 estimates of $\phi_1 = \mathrm{Cor}(X_{t,1}, X_{t-1,1})$ and $\phi_2 = \mathrm{Cor}(X_{t,2}, X_{t-1,2})$ are very close. We thus have substantial evidence that the additional structure of Model 3 is unnecessary for the EV Lac data, and we proceed with an analysis of Model 2.

6.2 Stage 2: flaring/quiescent interval estimates for EV Lac

In this section, we demonstrate our Stage 2 methods for classifying the light curve, $\boldsymbol{Y}_{1:T}$, into flaring and quiescent intervals by fitting finite mixture distributions to the predicted states $\hat{X}_1, \dots, \hat{X}_T$. All calculations in this section are under the preferred Model 2.

6.2.1 Semisupervised classification for ObsID 01885

The predicted state variables, given by

$$\hat{X}_t = \underset{x \in \mathcal{X}}{\operatorname{argmax}}\, \mathbb{P}_{\hat{\eta}}\left(X_t = x \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}\right), \quad t = 1, \dots, T \tag{31}$$

with $T = 2027$ are computed using the local decoding procedure described in Appendix A and plotted for ObsID 01185 in Fig. 5.

A visual inspection of the ObsID 01185 light curve in Fig. 1 and its predicted states in Fig. 5 reveals a clear period of quiescent equilibrium over the first 750 time bins. Thus, we could apply the semisupervised approach of Section 5.1 to model the distribution of the $\hat{X}_t$. After fitting the KDE $\hat{f}_1$ to $\{\hat{X}_1, \dots, \hat{X}_{750}\}$, we chose $K = 25$ ‘steps’ for the step function in equation (27), setting the intervals $[b_{k-1}, b_k)$ to be 25 evenly spaced subintervals in $[b_0, b_K]$, where $b_0 \approx -0.35$ is the median of $\hat{f}_1$ and $b_K = \sup A = 2.65$. (We assume that the lowest levels of activity correspond to quiescence.)

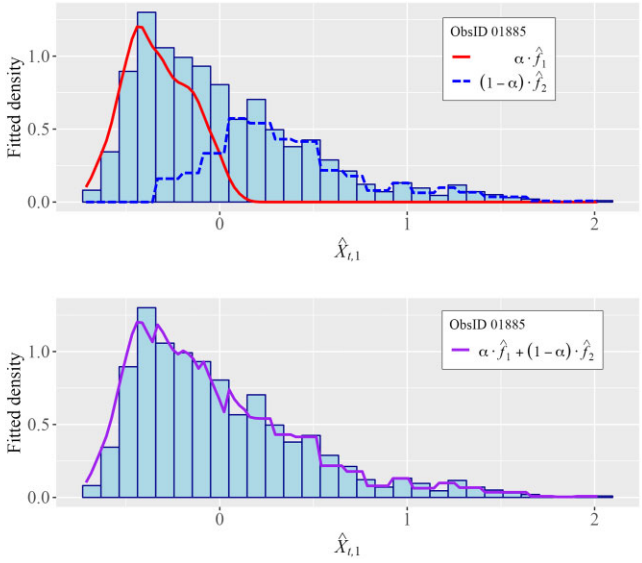

Figure 6. Fitted component densities (top panel) and mixture density (bottom panel) for ObsID 01885; the densities are overlaid on a histogram of $\{\hat{X}_1, \dots \hat{X}_{2027}\}$.

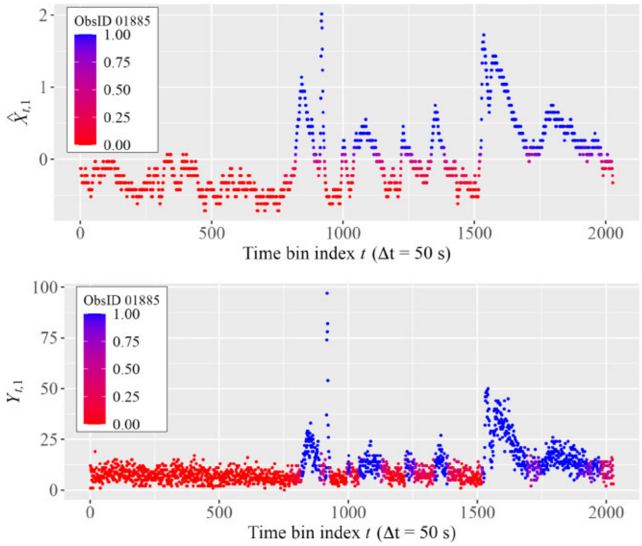

Figure 7. Posterior flaring state probabilities used to colour the predicted states $\hat{X}_1, \dots, \hat{X}_t$ (top panel) and the observed soft-band counts $Y_{1,1}, \dots, Y_{T,1}$ (bottom panel) for ObsID 01885.

We fit the mixture in equation (25) to $\{\hat{X}_{751}, \dots, \hat{X}_{2027}\}$ using the EM algorithm described in Appendix C1, which yielded a mixing parameter estimate $\hat{\alpha} = 0.5528$, indicating that the estimated proportion of time that EV Lac spends in a flaring state based on the ObsID 01885 time bin is $1 - \hat{\alpha} \approx 0.45$. The resulting component densities and mixture density are illustrated in Fig. 6.

Using equation (28), we computed the posterior flaring state probability for each $\hat{X}_t$; these are shown on a colour gradient in Fig. 7, both for the predicted states and the original soft-band counts, $Y_{1,1:T}$ (Fig. D1 in Appendix D shows the soft and hard band counts coloured by the same probabilities.) From the posterior flaring state probabilities, we created binary quiescent/flaring classifications $\hat{z}_1, \dots, \hat{z}_T \in \{1, 2\}$ using a simple classification rule, which is the basis for the results given in Section 6.2.3 below: letting

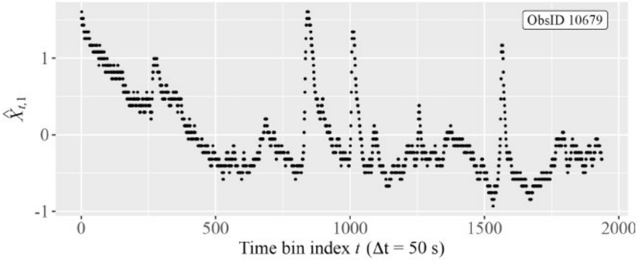

Figure 8. Predicted soft-band states $\hat{X}_1, \ldots \hat{X}_{2027}$ for ObsID 10679.

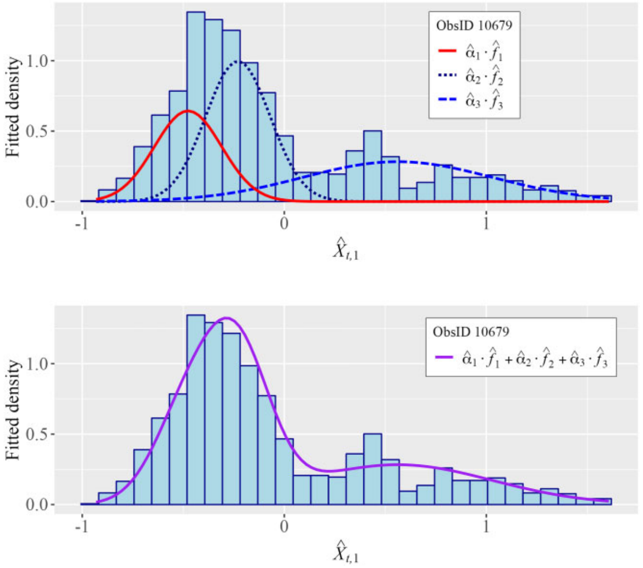

Figure 9. Fitted component densities (above) and mixture density (below) for ObsID 10697; the densities are overlaid on a histogram of $\{\hat{X}_1, \ldots \hat{X}_{1937}\}$.

$\hat{p}_t = \mathbb{P}(Z_t = 2 \mid \hat{X}_t = \hat{x}_t)$ as in equation (28), EV Lac was classified as being in a flaring state at time index t if and only if $\hat{p}_t > 0.5$; equivalently

$$\hat{z}_t = 1 \cdot \mathbb{1}_{\hat{p}_t \leq 0.5} + 2 \cdot \mathbb{1}_{\hat{p}_t > 0.5}. \tag{32}$$

6.2.2 *Unsupervised classification for ObsID 10679*

For the light curves in ObsID 10679, there is no clear period of quiescence (see the lower panel of Fig. 1). The soft-band predicted state variables, $\hat{X}_1, \ldots, \hat{X}_{1937}$, under Model 2 are plotted in Fig. 8, again illustrating the lack of a clearly sustained period of quiescence. Thus, we used this data to demonstrate the unsupervised classification method described in Section 5.2, fitting a mixture of three normal distributions to the complete set of predicted state variables.

The estimated parameters of the mixture components are given in Table 6 and the estimated component densities and mixture density are shown in Fig. 9. Under the assumption that the third component corresponds strictly to the flaring state, the estimated proportion of time that EV Lac spends in a flaring state based on the ObsID 10 679 data is $\hat{\alpha}_3 \approx 0.27$.[8] Corresponding posterior flare probabilities, which associate the third component of the mixture model (29) to the flaring state, are shown in Fig. 10 (and in Fig. D1 for counts in both bands). A binary classification rule nearly identical to that described in Section 6.2.1 was created, the only difference being that now $\hat{p}_t = \mathbb{P}(Z_t = 3 \mid \hat{X}_t = \hat{x}_t)$; binary quiescent/flaring classifications were again constructed according to (32).

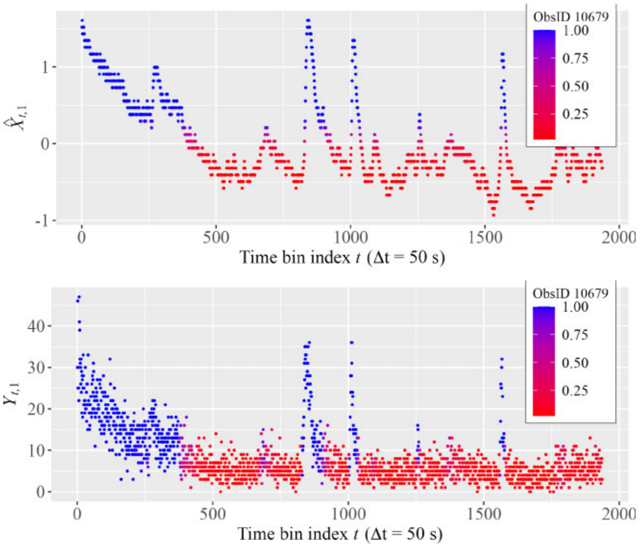

Figure 10. Posterior flaring state probabilities used to colour the predicted states $\hat{X}_1, \ldots, \hat{X}_t$ (above) and the observed soft-band data $Y_{1,1}, \ldots, Y_{T,1}$ (below) for ObsID 10697.

6.2.3 *The quiescent and flaring states of EV Lac*

We also carried out sensitivity checks on the flaring intervals determined as above in Sections 6.2.1 and 6.2.2 by jittering the phase of the binning by ±25 s, changing the passband intervals (using 0.3–0.9 keV for the softer and 0.9–8.0 keV for the harder bands), and checking different time bin widths (see Appendix D). We found that the flaring intervals thus determined remain stable and repeatable to within 2–3 time bin widths in all cases. We thus adopted a 3× time bin width as a nominal systematic uncertainty on the intervals, and merged all gaps smaller than that. We further inflated the intervals by adding 25 s (half the width w of the adopted time bins) both before and after the ends of each interval. This resulted in 15 distinct intervals for ObsID 01885 and 11 intervals for ObsID 10 679 (see Tables 7 and 8, respectively). The durations of the interval correspond to approximately 30 per cent and 40 per cent of the total observation interval for the first and second epochs, respectively. This is consistent with the expected flare rates seen on EV Lac before: flares occurring at rates of 0.2–0.4 h^{-1} (Huenemoerder et al. 2010) lasting approximately 5 ks cover a fraction of 0.28–0.55 of

[8] We associate the first two components of the mixture distribution with quiescence because the fitted density shown in Fig. 9 indicates considerable overlap between these two components. Alternatively, one could postulate that the second component is composed of both flaring and quiescent states and/or corresponds to the transition between the two states. Thus, $\hat{\alpha}_3$ may slightly underestimate the proportion of time that EV Lac spends in its flaring state. Of course, the mixture model is completely agnostic to our own astrophysical interpretations of its components. Possibly, both the second and the third components together correspond to a flaring state; under this assumption, the estimated proportion of time spent in this state is $\hat{\alpha}_2 + \hat{\alpha}_3 \approx 0.60$. In general, the interpretation of the distinction between quiescent and flaring states must be done on a case by case basis, as it can depend on the source, the epoch of observation, and the instrument being used. The spectral variability analysis presented in Section 6.2.3 strongly supports our interpretation in this case.

Table 6. Parameter estimates for the three-component mixture of normal distributions.

Component k	$\hat{\alpha}_k$	$\hat{\mu}_k$	$\hat{\tau}_k^2$
1	0.3988	−0.2294	0.0255
2	0.3328	0.5608	0.2202
3	0.2683	−0.4764	0.0277

Table 7. Flaring time intervals for ObsID 1885, in spacecraft clock time. The times are offset from the observation start time of 117315383.3 s, corresponding to a calendar time of 2001 September 19, 19:36:23.

Interval	Duration [s]	Start time [s]	Stop time [s]
1	4000	41624.1	45624.1
2	950	46224.1	47174.1
3	700	50474.1	51174.1
4	4900	52724.1	57624.1
5	100	58124.1	58224.1
6	2100	61774.1	63874.1
7	100	64474.1	64574.1
8	150	65324.1	65474.1
9	100	66874.1	66974.1
10	100	67174.1	67274.1
11	3000	67474.1	70474.1
12	300	71724.1	72024.1
13	23 250	76674.1	99924.1
14	600	100724.1	101324.1
15	100	101524.1	101624.1

Table 8. Flaring time intervals for ObsID 10679, in spacecraft clock time. The times are offset from the observation start time of 353314077.3 s, corresponding to a calendar time of 2009 March 13, 06:47:57.

Interval	Duration [s]	Start time [s]	Stop time [s]
1	19 850	1742.7	21592.7
2	250	21792.7	22042.7
3	150	22742.7	22892.7
4	600	35792.7	36392.7
5	4300	43192.7	47492.7
6	100	47892.7	47992.7
7	1800	51842.7	53642.7
8	150	56142.7	56292.7
9	450	64392.7	64842.7
10	100	70392.7	70492.7
11	1000	79692.7	80692.7

the exposure durations, assuming no overlaps. Note that our method does not distinguish the number of flares within a flare state (e.g. the first interval in ObsID 10 679 covers a duration that clearly includes a smaller flare that overlaps another with a longer decay time-scale).

This separation between flaring and quiescent states allows us to explore changes in the energy spectrum of the star. The overall spectrum is well fitted with a two-temperature component XSAPEC model in CIAO/SHERPA v4.16 (Refsdal et al. 2009) with similar temperature, abundance, and normalizations for both epochs (see Table 9).

Fig. 11 shows the changes in spectral colour for each of the flare intervals (marked in blue) compared to the combined quiescent interval (marked in red); all error bars were computed using Bayesian Estimation of Hardness Ratios (BEHR; Park et al. 2006). The colours were computed as log-ratios of counts in the soft (S: 0.3–0.9 keV) to medium (M: 0.9–2.0 keV) and medium to hard (H: 2.0–8.0 keV) bands. It is clear that all of the flaring intervals have harder spectra than the quiescent spectrum. The underlying grid, constructed for a two-temperature APEC model as for the full spectra (see Table 9) but with varying normalization and temperature for the high-temperature component, also demonstrates this quantitatively. The flaring intervals include the low-temperature component because the flares are likely confined to small regions in the corona, so that the quiescent corona continues to contribute to the emission, even as the emission is dominated by the flare. Note that the grids shift leftwards from the earlier epoch to the later, which is a consequence of the increased contamination buildup on the ACIS detector which reduces the soft effective area.

Finally, we show in Fig. 12 the full resolution combined HEG + MEG first-order spectra separately for the quiescent (upper panels) and flaring states (lower panels). Spectra from both epochs are overplotted, and deviations where the counts from one epoch exceed the other are marked in different shades. As is expected from the evolution in the soft effective area, the earlier epochs have systematically higher counts at longer wavelengths. The spectra are dominated by several prominent lines, such as those from Ne X (12.15 Å), Fe XVII (15.01 and 17.05 Å), and O VIII (18.96 Å) (see middle panels of Fig. 12). The density- and temperature-sensitive He-like O VII triplet (21.6, 21.8, and 22.1 Å of the resonance, intercombination, and forbidden lines) is visible in the right panels; higher density plasma is present in the flaring state, as shown by the higher ratio of the intercombination to forbidden lines. In the left panels, several high-temperature lines appear during the flaring state at short wavelengths (Ar XVIII 2.92 Å, Ar XVII 3.95 Å, S XVI 4.73 Å, and S XV 5.0 Å). The ratios of the temperature sensitive resonance lines of Si XIV (6.2 Å) and Si XIII (6.74 Å), and Mg XII (8.4 Å) and Mg XI (9.2 Å) change to favour the higher temperature species and the continuum becomes more prominent, all indicating the presence of higher temperature plasma, and thus supporting the conclusions of Huenemoerder et al. (2010).

In addition, the Ne X/O VIII counts ratio increases from 2.1 during quiescence to 3.5 ± 0.2 in the first epoch, and from 2.6 to 3.4 ± 0.3 during the second epoch. The Ne X/Fe XVII counts ratio also increases, from approximately 2.8 during quiescence to approximately 3.5–4.0 during flaring in both epochs, indicating that there could be an increase in Ne abundance during flaring. In contrast, the O VIII/Fe XVII counts ratio decreases by approximately 10 per cent during flaring in both epochs; detailed modelling is necessary to establish whether this decrease is simply a temperature effect or whether oxygen abundance variations are also required to explain it.

Crucially, the differences between epochs for each state are minuscule compared to the changes seen between the quiescent and flaring states. This is a strong indication that our method can clearly identify and separate these states. Furthermore, the similarity in the apparent thermal characteristics in both states, as evidenced by the similar shapes of the continuum, shows that the two states are strongly differentiated: that is, the star has a very well-defined quiescent state, suggesting that there may be a distinct heating mechanism that operates during quiescence.

7 DISCUSSION AND FUTURE WORK

This paper combines state-space models and finite mixture models as a means of classifying periods of quiescence and flaring in multiband astronomical light curves. Specifically, we apply our models to high-energy X-ray data of the active binary EV Lac, grouping the photons into two passbands and classifying the light curves into flaring and quiescent states. In Stage 1 of our analysis, our state-space models

Table 9. SHERPA two-temperature APEC model fits to the full spectra.

ObsID	$T_{\rm low}$	$T_{\rm high}$	Metallicity	Norm$_{\rm low}$	Norm$_{\rm high}$	CSTAT/dof
	[keV]		$Z_\odot$	[$\times 10^{14}$ cm^{-5}]		
01885	0.35±0.0024	1.26±0.007	0.17±0.004	0.016±0.0004	0.0099±0.0001	24850.1/24980
10 679	0.35±0.003	1.35±0.009	0.17±0.005	0.015±0.0005	0.0095±0.0001	21668.7.1/24980

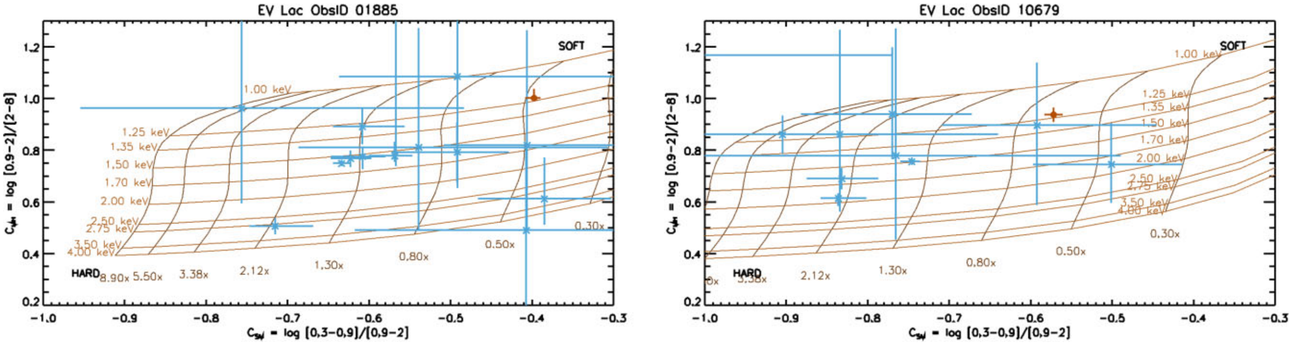

Figure 11. Spectro-temporal treatment of flaring. For both EV Lac data sets, the panels show the hardness ratio colours $C_{SM} = \log(S/M)$, $C_{MH} = \log(M/H)$, where S, M, and H are counts in passbands 0.3–0.9, 0.9–1.2, and 1.2–8.0 keV, respectively. The colours during each distinct flaring interval (crosses with error bars) are compared to the quiescent interval (sole round point with error bars). The curved grid in the background shows the predicted colours for spectra with two temperature components: a low-temperature plasma at $T_{\rm low} = 0.35$ keV (≈4 MK), and a high-temperature component with a variety of temperatures $T_{\rm high}$ ranging from 1 keV (≈12 MK) to 4 keV (≈46 MK), with the relative emission measure of the high-temperature component ranging between 0.1 to 8.9 times that of the low-temperature component. We adopt a metallicity of 0.16, commensurate with a two-temperature APEC fit to the spectra. Note that in both epochs, the quiescent interval has a softer spectrum than any of the flaring intervals. The shift in the grid is due to changes in ACIS effective area between the epochs.

(HMMs) assume that the underlying physical process driving the flaring activity can be represented by a Markov chain defined on a continuous multidimensional state space. When the component of the Markov chain corresponding to a particular energy band migrates to higher or lower values, the rate of photon emissions within that band tends to increase or decrease in kind. We propose a series of nested HMMs to capture this underlying process with increasing levels of generality. We tabulate emissions in the soft and hard energy bands separately in order to capture the more complete information contained in the bivariate data. The state-space models allow us to predict the individual states of the underlying chain that are most likely to have generated the observed data. Using finite mixture models in Stage 2, we devise two situation-specific schemes to classify the predictions and ultimately dichotomize the observed time periods into flaring and quiescent intervals.

7.1 Quiescence

We demonstrate our method on two sets of observations of the dMe star EV Lac, leading to a clear separation of flaring activity and quiescence, as well as to the discovery of a well-defined and persistent quiescent state. The presence of such persistent quiescent emission in counterpoint to flaring has been recognized and analysed ubiquitously in astronomical literature. Exemplar treatments include that of the Sun by Argiroffi et al. (2008) and of an active M dwarf YZ CMi by Raassen, Mitra-Kraev & Güdel (2007); see also a review by Güdel (2004). The possibility of a persistent quiescent state has also been suggested for the active binary AR Lac (Drake et al. 2014), and for the young stellar binary XZ Tau (Silverberg et al. 2023). Our analysis of spectral variability supports the idea that steady and persistent non-flaring emission is present even on active stars.

The continuous-space HMMs that we propose in our Stage 1 analysis (Section 3) do not alone clearly differentiate between the quiescent and active states of the source, instead allowing for variability within the states and a smooth transition between them. The time intervals during which flaring emission dominates are identified from the distribution of the fitted HMM states in our Stage 2 analysis (Section 5). Alternatively, we could posit a model where the quiescent emission is present at all times, with the intermittent and variable flaring emission (presumably arising in localized active regions on the star) superposed over it. For example, the observed counts could be modelled as the sum of two Poisson processes, the first an iid process representing quiescence and the second representing the flare state alone. Such a model would be more complex than the HMMs we consider here in that its second Poisson process (for the flaring state) would be as complex as the HMMs we propose in Section 4. As we expect that the flexibility of the continuous-space HMM may render the more complex model unidentifiable or only weakly identifiable, we leave its study for future research.

7.2 Future directions

We propose several avenues for future modifications and generalizations of our HMMs. The discrete-space HMM approximation to the state-space likelihood developed by Langrock (2011) is, in the end, only an approximation, which can potentially be made more accurate using adaptive binning (Borowska & King 2023) or other procedures that further refine the discretization of the continuum (i.e. the choice of essential domain and the partition thereof). From a computational perspective, it would also be desirable to eliminate the need for manual verification that the essential domain adequately covers the distribution of the underlying Markov chain.

The state-space models themselves can be modified or augmented with additional features. In Section 4.2, for example, we discuss the use of state-dependent bivariate distributions for the observed data. In the general case, this avoids a conditional independence assumption for the hard and soft energy bands, and allows for

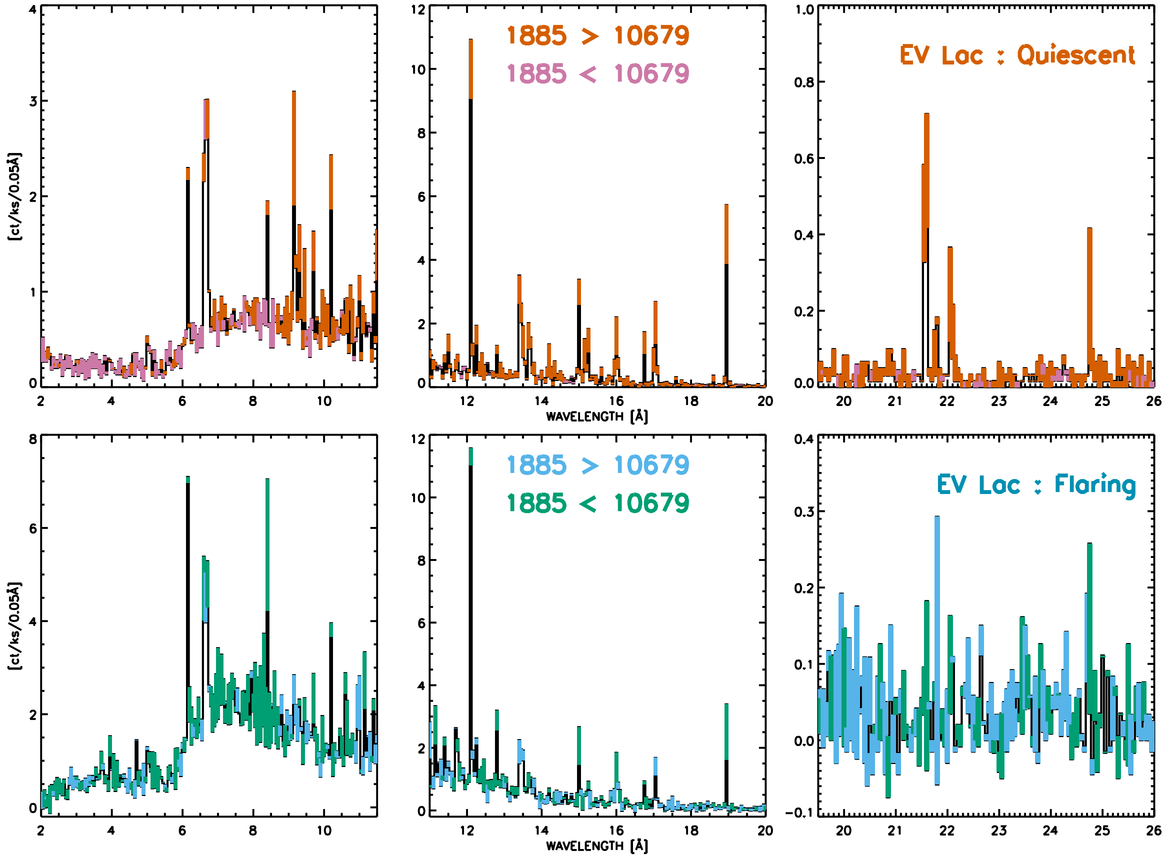

Figure 12. Comparing high-resolution spectra of quiescent (top) and flaring (bottom) states of active dMe star EV Lac. The quiescent spectra are subtracted out from the corresponding spectra obtained during the flaring state. Each state is split into three panels in order to better show weak lines. Spectra from the two epochs are shown superposed for both cases; the difference between the epochs is marked in different shades depending on which epoch had more counts within a given bin (see the ObsID labels in the middle panels). Prominent lines from several species are visible in both spectra, with some resonance line ratios changing from quiescence to a flaring state favouring the higher ionization state (see especially Si XIV/Si XIII at 6.2 and 6.7 Å, Mg XII/Mg XI at 8.4 and 9.2 Å). The overall brightness is higher, and the continuum is stronger and more prominent during the flaring state, signifying a different thermal signature. The differences between the quiescent and flaring states are greater than the differences between epochs for the same state, which suggests that there are distinct quiescent and flaring states present on the star.

more involved bivariate distributions capable of capturing potential dependence between the bands at the observed data level. Even more generally, one could split the counts into any number d of bands (the hard and soft bands we used for EV Lac correspond to $d = 2$). The d-band generalization of Model 2 is straightforward: for each additional band h, we introduce one new parameter β_h controlling the Poisson rate for $Y_{t,h}$, as well as a rescaling parameter σ_h so that $X_{t,h} = \sigma_h X_{t,1}/\sigma_1$. The generalization of Model 3 is more challenging due to the increased complexity of the (non-diagonal) covariance matrix $\boldsymbol{\Sigma}$ in the error terms: in addition to new parameters β_h, ϕ_h, and σ_h, each band h requires pairwise correlation terms with every other dimension, resulting in a $d \times d$ covariance matrix $\boldsymbol{\Sigma}$. Depending on the covariance structure selected for the model, $\boldsymbol{\Sigma}$ can include as few as two free parameters (for a first-order autoregressive covariance) or as many as $(d^2 + d)/2$ (for a completely unstructured covariance). For large d, this would effectively model the evolution of the spectrum over time. This model is in contrast to Automark, which looks for breakpoints in the spectrum but assumes the spectrum is unchanging between breakpoints (Wong et al. 2016).

It is also possible to generalize the distribution of the state process $\boldsymbol{X}_{1:T}$ by replacing the multivariate normal distributions with other multivariate distributions. For instance, one could account for potentially heavier tails in the distribution of $\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}$ by assuming a multivariate t-distribution; alternatively, one could assume that $\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}$ follows a mixture of conditional multivariate normal distributions with common mean $\boldsymbol{X}_{t-1}$ but differing variances, which could potentially model a discrete latent process taking place in some physical process within the star itself. Both of these generalize the multivariate normal distribution, and their associated stationary distributions are available (e.g. Meitz, Preve & Saikkonen 2023); however, stationary distributions corresponding to other choices of the distribution of $\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}$ may not be known, and therefore a different distribution would be needed for the initial state $\boldsymbol{X}_0$. The effect of this choice is likely small with large data sets.

Even when adhering to multivariate normal conditional distributions for the state process, our models can be generalized in several other ways. For example, the VAR(1) model (20) can be generalized to allow Φ to be a generic asymmetric non-diagonal matrix; in this case, stationarity is characterized by a rather complex set of nonlinear constraints on the entries of Φ. This generalization would allow for dependence of $X_{t,1}$ on $X_{t-1,2}$, and vice versa. Such dependencies can be used to capture physical processes where hot coronal plasma in a magnetic flux tube cools sequentially from higher to lower temperatures (e.g. Viall & Klimchuk 2012). One can also consider more general VAR(p) processes (i.e. where the distribution of $\boldsymbol{X}_t$ depends linearly on $\boldsymbol{X}_{t-1}, \dots, \boldsymbol{X}_{t-p}$). Any discrete-time stochastic process $(X_t)_t$ on a state-space $\mathcal{X}$ for which the distribution of X_t depends on the history of the chain through $X_{t-1}, \dots, X_{t-p}$ (a so-called higher-order Markov chain) induces a standard vector-valued Markov chain $(X'_t)_t$ on $\mathcal{X}^p$, and so, in principle, a VAR(p) process on $\mathcal{X}^d$ can be recast as a first-order matrix (or 'tensor') autoregressive process on $\mathcal{X}^{d\times p}$ for which the discrete-space HMM approximation can be applied. However, the calculations required for the initial distribution and transition probabilities would involve the so-called matrix normal distribution, which can be quite computationally involved.

Additionally, rather than binning the photon counts into discrete intervals, one could model the series of photon counts directly in continuous time. This would involve modelling the exponentially distributed waiting time between the Poisson process of photon arrivals. The underlying state process would evolve in continuous time and could be modelled as an Ornstein–Uhlenbeck (OU) process, the continuous-time analogue of the AR(1) process. The OU process has been applied in astrophysical settings by Kelly, Bechtold & Siemiginowska (2009), Kelly et al. (2014), and Meyer et al. (2023); such processes generalize fairly naturally to the multivariate case (Gardiner 2004). Perhaps the most natural continuous-time analogue of our state-space model is a bivariate time-heterogeneous Poisson process (Cox & Lewis 1972) whose parameters are driven by components of the aforementioned OU process.

Finally, our methods can also be generalized to apply to sources other than stars that also exhibit intermittent or episodic flaring (e.g. Sgr A$*$, the jet of M87, or dipping sources such as LMXBs). Such generalizations would require our underlying HMMs to be extended in order to model additional passbands and their possible correlations.

ACKNOWLEDGEMENTS

This work was conducted under the auspices of the CHASC International Astrostatistics Center. CHASC is supported by National Science Foundation grants DMS-21-13615, DMS-21-13397, and DMS-21-13605; by the UK Engineering and Physical Sciences Research Council [EP/W015080/1]; and by NASA grant 80-NSSC21-K0285. RZ was supported by a Mitacs Globalink Research Award. DvD was supported in part by a Marie-Skłodowska-Curie RISE (H2020-MSCA-RISE-2019-873089) Grant provided by the European Commission. VLK and AS acknowledge support from NASA Contract NAS8-03060 to the *Chandra X-ray Center*. We thank Yanbo Tang for helpful discussions.

DATA AVAILABILITY

This paper employs a list of *Chandra* data sets, obtained by the *Chandra X-ray Observatory*, contained in the *Chandra* Data Collection (CDC) 235, doi:10.25574/cdc.235.

Data were obtained and reduced using CIAO (Fruscione et al. 2006). Spectral fitting and parameter grid calculations were carried out using SHERPA (Freeman, Doe & Siemiginowska 2001; Refsdal et al. 2009). Hardness ratios were computed using BEHR (Park et al. 2006). Flux estimates and line identifications were carried out in PINTofALE (Kashyap & Drake 2000). All state-space modelling was conducted in R (R Core Team 2023) with the aid of several packages; all R code is publicly available at github.com/rob-zimmerman/SSM-flare.

REFERENCES

Argiroffi C., Peres G., Orlando S., Reale F., 2008, A&A, 488, 1069
Aschwanden M. J., 2019, ApJ, 887, 57
Aschwanden M. J., Freeland S. L., 2012, ApJ, 754, 112
Borowska A., King R., 2023, J. Comput. Graph. Stat., 32, 19
Canizares C. R. et al., 2005, PASP, 117, 1144
Cappé O., Moulines E., Ryden T., 2005, Inference in Hidden Markov Models. Springer Series in Statistics. Springer-Verlag, Berlin, Heidelberg
Cox D. R., Lewis P. A. W., 1972, in Le Cam L.M., Neyman J., Scott E. L., eds, Proc. Sixth Berkeley Symposium on Mathematical Statistics and Probability. Univ. California Press, Berkeley, California, p. 401
Davenport J. R. et al., 2014, ApJ, 797, 122
Davis J. E., 2001, ApJ, 562, 575
Dempster A. P., Laird N. M., Rubin D. B., 1977, J. R. Stat. Soc.: Ser. B, 39, 1
Drake J. J., Ratzlaff P., Kashyap V., Huenemoerder D. P., Wargelin B. J., Pease D. O., 2014, ApJ, 783, 2
van Dyk D. A., Meng X.-L., 2010, Stat. Sci., 25, 429
Efron B., Tibshirani R. J., 1993, An Introduction to the Bootstrap. Chapman and Hall/CRC Monographs on Statistics and Applied Probability, Chapman and Hall/CRC, London
Esquivel J. A., Shen Y., Leos-Barajas V., Eadie G., Speagle J., Craiu R. V., Medina A., Davenport J., 2024, preprint (arXiv:2404.13145)
Feinstein A. D., Montet B. T., Ansdell M., Nord B., Bean J. L., Günther M. N., Gully-Santiago M. A., Schlieder J. E., 2020, AJ, 160, 219
Fleming S. W., Million C., Osten R. A., Kolotkov D. Y., Brasseur C., 2022, ApJ, 928, 8
Freeman P., Doe S., Siemiginowska A., 2001, in Starck J.-L., Murtagh F. D., eds, Proc. SPIE Conv. Ser. Vol. 4477, Astronomical Data Analysis. SPIE, Bellingham, p. 76
Fruscione A. et al., 2006, in Silva D. R., Doxsey R. E., eds, Proc. SPIE Conf. Ser. Vol. 6270, Observatory Operations: Strategies, Processes, and Systems. SPIE, Bellingham, p. 62701V
Gardiner C. W., 2004, Handbook of Stochastic Methods: For Physics, Chemistry and the Natural Sciences, third edn, Springer Series in Synergetics, Vol. 13. Springer-Verlag, Berlin
Gerber F., Furrer R., 2019, The R Journal, 11, 352
Güdel M., 2004, A&ARv., 12, 71
Hamilton J. D., 2020, Time Series Analysis. Princeton Univ. Press, Princeton, NJ
Huenemoerder D. P., Schulz N. S., Testa P., Drake J. J., Osten R. A., Reale F., 2010, ApJ, 723, 1558
Johnson N., Kotz S., Balakrishnan N., 1997, Discrete Multivariate Distributions. Wiley Series in Probability and Statistics. Wiley, New York
Kashapova L. K., Broomhall A.-M., Larionova A. I., Kupriyanova E. G., Motyk I. D., 2021, MNRAS, 502, 3922
Kashyap V., Drake J. J., 2000, Bull. Astron. Soc. India, 28, 475
Kelly B. C., Bechtold J., Siemiginowska A., 2009, ApJ, 698, 895
Kelly B. C., Becker A. C., Sobolewska M., Siemiginowska A., Uttley P., 2014, ApJ, 788, 33
Kitagawa G., 1987, J. Am. Stat. Assoc., 82, 1032
Langrock R., 2011, J. Appl. Stat., 38, 2955
Langrock R., MacDonald I. L., Zucchini W., 2012a, J. Empir, Financ., 19, 147
Langrock R., King R., Matthiopoulos J., Thomas L., Fortin D., Morales J. M., 2012b, Ecology, 93, 2336
Lei W. H., Li C., Chen F., Zhong S. J., Xu Z. G., Chen P. F., 2020, MNRAS, 494, 975
Lepreti F., Carbone V., Veltri P., 2001, ApJ, 555, L133
McLachlan G. J., Krishnan T., 2007, The EM Algorithm and Extensions. Wiley Series in Probability and Statistics. John Wiley and Sons, New York
Meitz M., Preve D., Saikkonen P., 2023, Commun. Stat. – Theory Methods, 52, 499
Meyer L., Witzel G., Longstaff F., Ghez A., 2014, ApJ, 791, 24
Meyer A. D., van Dyk D. A., Tak H., Siemiginowska A., 2023, ApJ, 950, 37
Moon Y.-J., Choe G. S., Yun H. S., Park Y. D., 2001, J. Geophys. Res.: Space Phys., 106, 29951https://doi.org/10.1029/2000JA000224

Muirhead R. J., 2009, Aspects of Multivariate Statistical Theory. Wiley Series in Probability and Statistics. John Wiley and Sons, New York

Nalewajko K., 2013, MNRAS, 430, 1324

National Academies of Science, Engineering, and Medicine, 2021, Pathways to Discovery in Astronomy and Astrophysics for the 2020s. The National Academies Press, Washington, DC

Park T., Kashyap V. L., Siemiginowska A., van Dyk D. A., Zezas A., Heinke C., Wargelin B. J., 2006, ApJ, 652, 610

Peck C., Machol J., Codrescu S., Zetterlund E., Rachmeler L., Viereck R., 2021, in AGU Fall Meeting Abstracts. p. SH25E–2139

Plucinsky P. P., Bogdan A., Marshall H. L., 2022, in den Herder J.-W. A., Nikzad S., Nakazawa K., eds, Proc. SPIE Conf. Ser. Vol. 12181, Space Telescopes and Instrumentation 2022: Ultraviolet to Gamma Ray. SPIE, Bellingham, p. 121816X

Primiceri G. E., 2005, Rev. Econ. Stud., 72, 821

Protassov R., van Dyk D. A., Connors A., Kashyap V. L., Siemiginowska A., 2002, ApJ, 571, 545

R Core Team, 2023, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, available at: https://www.R-project.org/

Raassen A. J. J., Mitra-Kraev U., Güdel M., 2007, MNRAS, 379, 1075

Rabiner L. R., 1989, Proc. IEEE, 77, 257

Refsdal B. L. et al., 2009, in Proc. 8th Python in Science Conference. Texas, Austin, p. 51

Resnick S. I., 2013, Adventures in Stochastic Processes. Birkhäuser Boston, Boston, MA

Rivera E. C., Johnson J. R., Homan J., Wing S., 2022, ApJ, 937, L8

Robinson R., Carpenter K., Percival J., Bookbinder J., 1995, ApJ, 451, 795

Scargle J. D., 1998, ApJ, 504, 405

Scargle J. D., Norris J. P., Jackson B., Chiang J., 2013, ApJ, 764, 167

Self S. G., Liang K.-Y., 1987, J. Am. Stat. Assoc., 82, 605

Silverberg S. M., Günther H. M., Pradhan P., Principe D. A., Schneider P. C., Wolk S. J., 2023, AJ, 166, 148

Stanislavsky A., Nitka W., Małek M., Burnecki K., Janczura J., 2020, J. Atmos. Sol.-Terr. Phys., 208, 105407

Szudzik M., 2006, in Wolfram Research (ed.) Special NKS 2006 Wolfram Science Conference. p. 1

Viall N. M., Klimchuk J. A., 2012, ApJ, 753, 35

Weisskopf M. C., Brinkman B., Canizares C., Garmire G., Murray S., Van Speybroeck L. P., 2002, PASP, 114, 1

Wheatland M. S., 2000, ApJ, 536, L109

Wheatland M. S., Litvinenko Y. E., 2002, Sol. Phys., 211, 2550

Wong R. K. W., Kashyap V. L., Lee T. C. M., van Dyk D. A., 2016, Ann. Appl. Stat., 10, 1107

Xu C., Günther H. M., Kashyap V. L., Lee T. C., Zezas A., 2021, AJ, 161, 184

Yoshida K., Petropoulou M., Murase K., Oikonomou F., 2023, ApJ, 954, 194

Zimmerman R., Craiu R. V., Leos-Barajas V., 2023, J. Am. Stat. Assoc., 1

Zucchini W., MacDonald I. L., Langrock R., 2017, Hidden Markov Models for Time Series: An Introduction using R, 2nd edn. Chapman and Hall/CRC, New York, NY

APPENDIX A: ALGORITHMS FOR DISCRETE-SPACE HMMS

Likelihood computation and state decoding for discrete-space HMMs generally rely on several related algorithms, including the forward algorithm, the backward algorithm, and the forward–backward algorithm. Here, we briefly derive these three algorithms, each of which can be succinctly described by an iterated sequence of matrix multiplications; each step of the algorithms thus involves multiplying the matrix obtained in the previous step by a new matrix, yielding in the end a product of matrices which we show to be equivalent to a quantity of interest, such as the value of the likelihood function. The algorithms use dynamic programming to efficiently compute certain quantities; for example, a naïve computation of equation (2) via direct summation would require a number of operations exponential in T, while the forward algorithm reduces the computation to being only polynomial in T. In the HMM literature (e.g. Zucchini et al. 2017), these algorithms are typically expressed in terms of quantities known as forward and backward variables, but these are unnecessary for our purposes as the relevant equations can be succinctly expressed in terms of matrices alone. Nevertheless, the forward and backward variables play key roles in the theory of HMMs; interested readers may consult Rabiner (1989), Cappé et al. (2005), and Zucchini et al. (2017).

To simplify notation, we assume that the $\boldsymbol{y}_t$ are discrete in our derivations, as is the case in the Poisson models developed in Section 4. None the less, our calculations carry through verbatim for continuous observations, with probability mass functions replaced by their analogous density functions.

A1 Likelihood computation via the forward algorithm

The forward algorithm for discrete-space HMMs evaluates the HMM likelihood $L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T})$ given by equation (2) via an efficient computation of the right-hand side of the identity

$$L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}) = \mathbb{P}_{\boldsymbol{\eta}}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}). \tag{A1}$$

For the remainder of this section, we drop the subscript $\boldsymbol{\eta}$ from $\mathbb{P}_{\boldsymbol{\eta}}(\cdot)$ for notational simplicity; however, all probabilities should be understood as being taken with respect to the model with parameter $\boldsymbol{\eta}$.

Define the matrix-valued function $\boldsymbol{P} : \mathcal{Y} \to [0, \infty)^{K \times K}$ by

$$\boldsymbol{P}(\boldsymbol{y}) = \begin{bmatrix} h_1(\boldsymbol{y} \mid \boldsymbol{\lambda}_1) & 0 & \cdots & 0 \\ 0 & h_2(\boldsymbol{y} \mid \boldsymbol{\lambda}_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & h_K(\boldsymbol{y} \mid \boldsymbol{\lambda}_K) \end{bmatrix}; \tag{A2}$$

that is, $\boldsymbol{P}(\boldsymbol{y})$ is a diagonal matrix with the state-dependent mass/density functions $h_1, \ldots, h_K$ evaluated at $\boldsymbol{y}$ along its diagonal. Recalling that $h_k(\boldsymbol{y} \mid \boldsymbol{\lambda}_k) = \mathbb{P}(\boldsymbol{Y}_t = \boldsymbol{y} \mid X_t = k)$ for any t we have

$$\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1) = \begin{bmatrix} \mathbb{P}(X_1 = 1) \cdot \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1 \mid X_1 = 1) \\ \vdots \\ \mathbb{P}(X_1 = K) \cdot \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1 \mid X_1 = K) \end{bmatrix}^\top = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, X_1 = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, X_1 = K) \end{bmatrix}^\top \tag{A3}$$

and

$$\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1)\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_2) = \begin{bmatrix} \sum_{k=1}^{K} \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, X_1 = k)\cdot \mathbb{P}(X_2 = k \mid X_1 = 1)\cdot \mathbb{P}(\boldsymbol{Y}_2 = \boldsymbol{y}_2 \mid X_2 = k) \\ \vdots \\ \sum_{k=1}^{K} \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, X_1 = k)\cdot \mathbb{P}(X_2 = k \mid X_1 = K)\cdot \mathbb{P}(\boldsymbol{Y}_2 = \boldsymbol{y}_2 \mid X_2 = k) \end{bmatrix}^\top$$

$$= \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, X_2 = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, X_2 = K) \end{bmatrix}^\top. \tag{A4}$$

The forward algorithm iterates this matrix multiplication, and by induction at iteration $t \geq 2$, the algorithm returns

$$\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1)\prod_{s=2}^{t}\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_s) = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \cdots, \boldsymbol{Y}_t = \boldsymbol{y}_t, X_t = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_1 = \boldsymbol{y}_1, \boldsymbol{Y}_2 = \boldsymbol{y}_2, \cdots, \boldsymbol{Y}_t = \boldsymbol{y}_t, X_t = K) \end{bmatrix}^\top = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = K) \end{bmatrix}^\top. \tag{A5}$$

Finally, the likelihood in equation (2) can be computed as

$$\left(\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1)\prod_{s=2}^{T}\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_s)\right)\mathbf{1} = \sum_{k=1}^{K}\mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}, X_T = k) = \mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) = L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}). \tag{A6}$$

Thus, the forward algorithm computes the likelihood via the left-hand side of equation (A6). (In practice, one must usually rescale the probabilities with each additional matrix product to prevent numerical underflow.) This computation has a time complexity of $O(TK^2)$. Note that as a byproduct of the likelihood computations, the forward algorithm also yields the joint probabilities $\mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = k)$ for any $t \in \{1, \ldots, T\}$, which are used in the local decoding algorithm (see Appendix A2).

A2 Local decoding via the forward–backward algorithm

The forward–backward algorithm for discrete-space HMMs evaluates the conditional state-membership probabilities given the full data set (i.e. $\mathbb{P}(X_t = k \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T})$, for each $k \in \mathcal{X}$ and each $t \geq 1$); these are then used to predict the state variables via equation (31), as we describe below. The forward–backward algorithm itself comprises of two subalgorithms – the forward algorithm, which computes the joint probabilities $\mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = k)$, and the backward algorithm, which computes the conditional probabilities $\mathbb{P}(\boldsymbol{Y}_{t:T} = \boldsymbol{y}_{t:T} \mid X_{t-1} = k)$ for each $t \geq 2$. A final combination of the forward and backward algorithms yields the desired conditional state-membership probabilities (i.e. $\mathbb{P}(X_t = k \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T})$). The forward algorithm, which also outputs the HMM likelihood $L(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}) = \mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T})$, is detailed in Appendix A1; we present the backward algorithm and the final combination step here.

Using the same notation as in Appendix A1, we first note that

$$(\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_T))\mathbf{1} = \begin{bmatrix} \sum_{k=1}^{K} \mathbb{P}(X_T = k \mid X_{T-1} = 1)\cdot \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_T = k) \\ \vdots \\ \sum_{k=1}^{K} \mathbb{P}(X_T = k \mid X_{T-1} = K)\cdot \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_T = k) \end{bmatrix} = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-1} = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-1} = K) \end{bmatrix} \tag{A7}$$

and

$$(\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_{T-1})\boldsymbol{\Gamma}\mathbf{P}(\boldsymbol{y}_T))\mathbf{1}$$

$$= \begin{bmatrix} \sum_{k=1}^{K} \mathbb{P}(X_{T-1} = k \mid X_{T-2} = 1)\cdot \mathbb{P}(\boldsymbol{Y}_{T-1} = \boldsymbol{y}_{T-1} \mid X_{T-1} = k)\cdot \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-1} = k) \\ \vdots \\ \sum_{k=1}^{K} \mathbb{P}(X_{T-1} = k \mid X_{T-2} = K)\cdot \mathbb{P}(\boldsymbol{Y}_{T-1} = \boldsymbol{y}_{T-1} \mid X_{T-1} = k)\cdot \mathbb{P}(\boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-1} = k) \end{bmatrix}$$

$$= \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_{T-1} = \boldsymbol{y}_{T-1}, \boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-2} = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_{T-1} = \boldsymbol{y}_{T-1}, \boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{T-2} = K) \end{bmatrix}. \tag{A8}$$

It then follows by induction that for any $t \in \{2, \ldots, T\}$,

$$\left(\prod_{s=t}^{T}\boldsymbol{\Gamma}\boldsymbol{P}(\boldsymbol{y}_s)\right)\mathbf{1} = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_t = \boldsymbol{y}_t, \ldots, \boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{t-1} = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_t = \boldsymbol{y}_t, \ldots, \boldsymbol{Y}_T = \boldsymbol{y}_T \mid X_{t-1} = K) \end{bmatrix} = \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_{t:T} = \boldsymbol{y}_{t:T} \mid X_{t-1} = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_{t:T} = \boldsymbol{y}_{t:T} \mid X_{t-1} = K) \end{bmatrix}. \tag{A9}$$

The backward algorithm computes the conditional probabilities $\mathbb{P}(\boldsymbol{Y}_{t:T} = \boldsymbol{y}_{t:T} \mid X_{t-1} = k)$ for each $t \geq 2$, via the left-hand side of equation (A9). The time complexity of this algorithm is also $O(TK^2)$.

With the quantities $\mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = k)$ and $\mathbb{P}(\boldsymbol{Y}_{t:T} = \boldsymbol{y}_{t:T} \mid X_{t-1} = k)$ in hand for each $t \in \{2, \ldots, T\}$, we have that for any $t \in \{2, \ldots, T-1\}$,

$$\begin{aligned}
&\frac{1}{\sum_{k=1}^{K} \mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}, X_T = k)} \begin{bmatrix} \mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = 1) \cdot \mathbb{P}(\boldsymbol{Y}_{(t+1):T} = \boldsymbol{y}_{(t+1):T} \mid X_t = 1) \\ \vdots \\ \mathbb{P}(\boldsymbol{Y}_{1:t} = \boldsymbol{y}_{1:t}, X_t = 1) \cdot \mathbb{P}(\boldsymbol{Y}_{(t+1):T} = \boldsymbol{y}_{(t+1):T} \mid X_t = K) \end{bmatrix} \\
&= \frac{1}{\mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T})} \begin{bmatrix} \mathbb{P}(X_t = 1) \cdot \mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T} \mid X_t = 1) \\ \vdots \\ \mathbb{P}(X_t = K) \cdot \mathbb{P}(\boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T} \mid X_t = K) \end{bmatrix} \\
&= \begin{bmatrix} \mathbb{P}(X_t = 1 \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) \\ \vdots \\ \mathbb{P}(X_t = K \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) \end{bmatrix},
\end{aligned} \tag{A10}$$

which is now a vector consisting of the desired conditional state-membership probabilities. Replacing the terms in the first expression of equation (A10) by equivalent quantities computed efficiently using the forward and backward algorithms, the overall forward–backward algorithm can itself be summarized concisely by the equivalent identity

$$\frac{1}{\left(\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1) \prod_{s=2}^{T} \boldsymbol{\Gamma} \boldsymbol{P}(\boldsymbol{y}_s)\right) \mathbf{1}} \left(\boldsymbol{\delta}^\top \boldsymbol{P}(\boldsymbol{y}_1) \prod_{s=2}^{t} \boldsymbol{\Gamma} \boldsymbol{P}(\boldsymbol{y}_s)\right)^\top \odot \left(\left(\prod_{s=t+1}^{T} \boldsymbol{\Gamma} \boldsymbol{P}(\boldsymbol{y}_s)\right) \mathbf{1}\right) = \begin{bmatrix} \mathbb{P}(X_t = 1 \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) \\ \vdots \\ \mathbb{P}(X_t = K \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) \end{bmatrix}, \tag{A11}$$

where $\odot$ refers to the element-wise (i.e. Hadamard) product of two matrices of equal dimension. The forward–backward algorithm refers to the computation of the conditional state-membership probabilities via one pass each of the forward and backward algorithms in order to compute equation (A10) for each $t \geq 2$. The time complexity of the forward–backward algorithm remains $O(TK^2)$.

After running the forward–backward algorithm, the local decoding procedure computes the most likely state of the Markov chain at each time index t given the observed data $\boldsymbol{Y}_{1:T}$ by simply selecting the coordinate corresponding to the largest entry in equation (A11). That is, we select

$$\hat{X}_t = \operatorname*{argmax}_{k \in \mathcal{X}} \mathbb{P}(X_t = k \mid \boldsymbol{Y}_{1:T} = \boldsymbol{y}_{1:T}) \tag{A12}$$

for each $t = 1, \ldots, T$, as required in equation (31).

APPENDIX B: LIKELIHOOD APPROXIMATION BY DISCRETE-SPACE HMMS

In this appendix, we show how the continuous-space HMM likelihood (3), which involves T iterated integrals over the state space $\mathcal{X}$, can be approximated by a quantity which is essentially of the form (2) and can be computed efficiently via the forward algorithm (see Appendix A1). Our presentation is based closely on the derivation of the univariate case in Langrock et al. (2012a), but applies to all three state-space models presented in Section 4.3, including the bivariate process described in Section 4.3.3. For generality, we present the approximation for an arbitrary continuous state space $\mathcal{X}$; for Models 1 and 2, $\mathcal{X} = \mathbb{R}$, and for Model 3, $\mathcal{X} = \mathbb{R}^2$. In the former case, each $\mathcal{X}$-valued vector below (e.g. $\boldsymbol{x}_t$, $\boldsymbol{c}_i^*$, etc.) is a univariate quantity.

The first step of the approximation is to identify an 'essential domain' $A \subset \mathcal{X}$ (Kitagawa 1987) for the $\boldsymbol{X}_t$, such that A is bounded and $\mathbb{P}(\boldsymbol{X}_t \notin A) = \mathbb{P}(\boldsymbol{X}_t \in A^c)$ is small for each t. We then partition A into a large number of subregions, $A_1, \ldots, A_m$; when $\mathcal{X} = \mathbb{R}$ it is convenient to use intervals and when $\mathcal{X} = \mathbb{R}^2$ we can use rectangles, possibly of different lengths and widths. We choose within each A_i a representative point $\boldsymbol{c}_i^*$, such as its centre. If the area of each A_i comprises a sufficiently small proportion of the total area of $A = \cup_{i=1}^{m} A_i$, then

$$\begin{aligned}
\int_{\mathcal{X}} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T &= \int_{A} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T + \int_{A^c} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T \\
&\approx \int_{A} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T \quad \text{since the integral over } A^c \text{ is assumed small} \\
&= \sum_{i_T=1}^{m} \int_{A_{i_T}} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T \\
&\approx \sum_{i_T=1}^{m} \int_{A_{i_T}} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \mathrm{d}\boldsymbol{x}_T \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right) \quad \text{since } \boldsymbol{x}_T \approx \boldsymbol{c}_{i_T}^* \text{ when } \boldsymbol{x}_T \in A_i \\
&= \sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \boldsymbol{X}_{T-1} = \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right).
\end{aligned} \tag{B1}$$

Thus, as $A \to \mathcal{X}$ and each $A_i \to \{\mathbf{c}_i^*\}$ (i.e. as the essential domain becomes larger and its partition becomes finer), the approximations above become more exact. Applying the same reasoning,

$$\int_{\mathcal{X}} \int_{\mathcal{X}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{x}_{T-1}}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{x}_{T-1}}\right) \cdot \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T \, \mathrm{d}\boldsymbol{x}_{T-1}$$

$$= \int_{\mathcal{X}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{x}_{T-1}}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{x}_{T-1}}\right) \cdot \left(\int_{\mathcal{X}} \gamma\left(\boldsymbol{x}_{T-1}, \boldsymbol{x}_T\right) \cdot h_{\boldsymbol{x}_T}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{x}_T}\right) \mathrm{d}\boldsymbol{x}_T\right) \mathrm{d}\boldsymbol{x}_{T-1}$$

$$\approx \int_{\mathcal{X}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{x}_{T-1}}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{x}_{T-1}}\right) \cdot \left(\sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \boldsymbol{X}_{T-1} = \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right) \mathrm{d}\boldsymbol{x}_{T-1}$$

approximating the inner integral by (B1)

$$\approx \int_{A} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{x}_{T-1}}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{x}_{T-1}}\right) \cdot \left(\sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \mathbf{X}_{T-1} = \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right) \mathrm{d}\mathbf{x}_{T-1}$$

since the integral over A^c is assumed small

$$= \sum_{i_{T-1}=1}^{m} \int_{A_{i_{T-1}}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{x}_{T-1}}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{x}_{T-1}}\right) \cdot \left(\sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \mathbf{X}_{T-1} = \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right) \mathrm{d}\boldsymbol{x}_{T-1}$$

$$\approx \sum_{i_{T-1}=1}^{m} \int_{A_{i_{T-1}}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \cdot h_{\boldsymbol{c}_{i_{T-1}}^*}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_{T-1}}^*}\right) \cdot \left(\sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \boldsymbol{X}_{T-1} = \boldsymbol{c}_{i_{T-1}}^*\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right) \mathrm{d}\boldsymbol{x}_{T-1}$$

since $\boldsymbol{x}_{T-1} \approx \boldsymbol{c}_{i_{T-1}}^*$ when $\boldsymbol{x}_{T-1} \in A_i$

$$= \sum_{i_{T-1}=1}^{m} \left[\int_{A_{i_{T-1}}} \gamma\left(\boldsymbol{x}_{T-2}, \boldsymbol{x}_{T-1}\right) \mathrm{d}\boldsymbol{x}_{T-1} \cdot h_{\boldsymbol{c}_{i_{T-1}}^*}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_{T-1}}^*}\right) \cdot \left(\sum_{i_T=1}^{m} \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \boldsymbol{X}_{T-1} = \boldsymbol{c}_{i_{T-1}}^*\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right)\right]$$

$$= \sum_{i_{T-1}=1}^{m} \sum_{i_T=1}^{m} \left(\mathbb{P}\left(\boldsymbol{X}_{T-1} \in A_{i_{T-1}} \mid \boldsymbol{X}_{T-2} = \mathbf{x}_{T-2}\right) \cdot h_{\boldsymbol{c}_{i_{T-1}}^*}\left(\boldsymbol{y}_{T-1} \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_{T-1}}^*}\right) \cdot \mathbb{P}\left(\boldsymbol{X}_T \in A_{i_T} \mid \boldsymbol{X}_{T-1} = \boldsymbol{c}_{i_{T-1}}^*\right) \cdot h_{\boldsymbol{c}_{i_T}^*}\left(\boldsymbol{y}_T \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_T}^*}\right)\right). \tag{B2}$$

Proceeding inductively and handling the edge case of $\boldsymbol{X}_1$ similarly, we obtain the approximation

$$L\left(\boldsymbol{\eta} \mid \boldsymbol{y}_{1:T}\right) \approx \sum_{i_1=1}^{m} \cdots \sum_{i_T=1}^{m} \left(\mathbb{P}\left(\boldsymbol{X}_1 \in A_{i_1}\right) \cdot h_{\boldsymbol{c}_{i_1}^*}\left(\boldsymbol{y}_1 \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_1}^*}\right) \prod_{t=2}^{T} \left(\mathbb{P}\left(\boldsymbol{X}_t \in A_{i_t} \mid \boldsymbol{X}_{t-1} = \boldsymbol{c}_{i_{t-1}}^*\right) \cdot h_{\boldsymbol{c}_{i_t}^*}\left(\boldsymbol{y}_t \mid \boldsymbol{\lambda}_{\boldsymbol{c}_{i_t}^*}\right)\right)\right), \tag{B3}$$

which is exactly equation (4).

APPENDIX C: EM ALGORITHMS

The EM algorithm (Dempster et al. 1977) is a popular tool used to fit statistical models in the presence of latent (or unobserved) data. Latent data may have a natural interpretation within the context of the problem (e.g. the $\boldsymbol{X}_{1:T}$ in equation (9a) representing the underlying physical process driving flaring activity is unobserved), or it may arise purely as a mathematical convenience to aid in inference (e.g. the Z_t in equation (26) representing component membership when a finite mixture distribution is used for non-parametric density estimation). The essential idea is to augment the observed data, $\boldsymbol{x}$, with 'missing' data, $\boldsymbol{Z}$, to form a complete data set, $(\boldsymbol{x}, \boldsymbol{Z})$, which induces a complete-data log-likelihood function $\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{Z})$. Similar to the ordinary log-likelihood function, $\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{Z})$ is simply the logarithm of the joint density of $(\boldsymbol{X}, \boldsymbol{Z})$, but viewed as a function of the underlying model parameter $\boldsymbol{\eta}$. The missing data, $\boldsymbol{Z}$, is user-selected and chosen to make $\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{x}, \boldsymbol{Z})$ more analytically tractable than the ordinary observed-data log-likelihood $\ell(\boldsymbol{\eta} \mid \boldsymbol{x})$. The EM algorithm is designed to compute the maximum-likelihood estimate – that is, the value of $\boldsymbol{\eta}$ that maximizes $\ell(\boldsymbol{\eta} \mid \boldsymbol{x})$ – by iteratively maximizing the conditional expectation of $\ell_{\text{com}}(\boldsymbol{\eta} \mid \mathbf{x}, \mathbf{Z})$, conditioned on the observed data $\boldsymbol{x}$. More formally, given a starting value of the parameter $\boldsymbol{\eta}^{(0)}$, the algorithm iterates between the following two steps,

E-step: Compute $Q(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(r)}) = \mathbb{E}_{\boldsymbol{\eta}^{(r)}}\left[\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{X}, \boldsymbol{Z}) \mid \boldsymbol{X} = \boldsymbol{x}\right]$
M-step: Set $\boldsymbol{\eta}^{(r+1)} = \underset{\boldsymbol{\eta}}{\operatorname{argmax}}\, Q\left(\boldsymbol{\eta} \mid \boldsymbol{\eta}^{(r)}\right)$

for $r = 1, 2, \ldots$ until convergence is achieved. In practice, if the equations in the E- and M-step admit closed-form solutions, the resulting algorithm can be specified as a set of recursive updates for the components of $\boldsymbol{\eta}^{(r+1)}$ in terms of $\boldsymbol{\eta}^{(r)}$.

The convergence properties of the EM algorithm have been well studied. One primary benefit of the EM algorithm is that at each step of the algorithm, the maximizing value produced by the M-step can never decrease the observed-data log-likelihood $\ell(\cdot \mid \boldsymbol{x})$ from its value at the previous iteration. Thus, the EM algorithm can only converge to a stationary point of the likelihood function (assuming such a point exists), and under broad regularity conditions this guarantees convergence to the MLE when the likelihood is unimodal. There is a rich literature on the EM algorithm and the numerous algorithms related to it; for more information, we refer the reader to the seminal paper by Dempster et al. (1977), the monograph by McLachlan & Krishnan (2007), and the review paper by van Dyk & Meng (2010).

In the following subsections, we briefly derive the EM algorithms used to fit the finite mixture models described in Section 6.2.

C1 For semisupervised classification

Here, the observed data $X_{1:T} = (X_1, \dots, X_T)$ is assumed to be an independent and identically distributed sample from the mixture distribution

$$F = \alpha \cdot F_1 + (1-\alpha) \cdot F_2(\cdot; \boldsymbol{\pi}), \tag{C1}$$

where F_1 is a known distribution and $F_2(\cdot; \boldsymbol{\pi})$ is a distribution with density

$$f_2(x; \boldsymbol{\pi}) = \sum_{k=1}^{K} \frac{\pi_k}{b_k - b_{k-1}} \cdot \mathbb{1}_{x \in [b_{k-1}, b_k)} = \prod_{k=1}^{K} \left(\frac{\pi_k}{b_k - b_{k-1}} \right)^{\mathbb{1}_{x \in [b_{k-1}, b_k)}}. \tag{C2}$$

Here $b_0, \cdots, b_K$ are known with $b_0 < b_1 < \cdots < b_K$, while $\alpha \in (0, 1)$ and $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ are parameters to be estimated, with $\sum_{k=1}^{K} \pi_k = 1$ and $\pi_k \in (0, 1)$ for each $k = 1, \dots, K$. We define the independent and identically distributed latent variables, $Z_{1:T} = (Z_1, \dots, Z_T)$, such that $Z_t \sim \mathrm{Bernoulli}(\alpha)$ and $X_t \mid (Z_t = k-1) \sim F_k$, for $k = 1, 2$. It is easy to verify that equation (C1) gives the marginal distribution of the X_t. The complete-data log-likelihood is

$$\begin{aligned} \ell_{\mathrm{com}}(\boldsymbol{\pi}, \alpha \mid x_{1:T}, Z_{1:T}) &= \log \left(\prod_{t=1}^{T} (\alpha \cdot f_1(x_t))^{Z_t} \cdot ((1-\alpha) \cdot f_2(x_t; \boldsymbol{\pi}))^{1-Z_i} \right) \\ &= \sum_{t=1}^{T} \left\{ Z_t \cdot \left[\log \left(\frac{\alpha}{1-\alpha} \right) + \log(f_1(x_t)) - \log(f_2(x_t; \boldsymbol{\pi})) \right] + \log(1-\alpha) + \log(f_2(x_t; \boldsymbol{\pi})) \right\}, \end{aligned} \tag{C3}$$

where

$$\log(f_2(x_t; \boldsymbol{\pi})) = \sum_{k=1}^{K} (\log(\pi_k) - \log(b_k - b_{k-1})) \cdot \mathbb{1}_{x_t \in [b_{k-1}, b_k)} \tag{C4}$$

The E-step requires the computation of $\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[\ell_{\mathrm{com}}(\boldsymbol{\pi}, \alpha; X_{1:T}, Z_{1:T}) \mid X_{1:T} = x_{1:T}]$, which by linearity requires only $\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[Z_t \mid X_{1:T} = x_{1:T}] = \mathbb{P}_{\boldsymbol{\eta}^{(r)}}(Z_t = 1 \mid X_t = x_t)$. Using Bayes' rule and the law of total probability, we find that

$$\mathbb{P}_{\boldsymbol{\eta}^{(r)}}(Z_t = 1 \mid X_t = x_t) = \frac{\alpha^{(r)} \cdot f_1(x_t)}{\alpha^{(r)} \cdot f_1(x_t) + \left(1 - \alpha^{(r)}\right) \cdot f_2\left(x_t; \boldsymbol{\pi}^{(r)}\right)} =: \gamma_1\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right). \tag{C5}$$

The M-step requires that we maximize

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[\ell_{\mathrm{com}}(\boldsymbol{\pi}, \alpha; X_{1:T}, Z_{1:T}) \mid X_{1:T} = x_{1:T}] \\ &= \sum_{t=1}^{T} \left\{ \gamma_1\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \left[\log \left(\frac{\alpha}{1-\alpha} \right) + \log(f_1(x_t)) - \log(f_2(x_t; \boldsymbol{\pi})) \right] + \log(1-\alpha) + \log(f_2(x_t; \boldsymbol{\pi})) \right\} \end{aligned} \tag{C6}$$

with respect to both α and $\boldsymbol{\pi}$; these optimizations can be carried out separately because these parameters are functionally independent in equation (C6). Basic calculus shows that the maximizing value of α is

$$\hat{\alpha} = \frac{1}{T} \sum_{t=1}^{T} \gamma_1\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right). \tag{C7}$$

Optimizing $\boldsymbol{\pi}$ is only slightly more complicated due to the constraint $\sum_{k=1}^{K} \pi_k = 1$, for which the method of Lagrange multipliers is particularly suitable. Applying this technique shows that π_k is maximized by

$$\hat{\pi}_k = \frac{\sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{x_t \in [b_{k-1}, b_k]}}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{x_t \in [b_{l-1}, b_l]}}, \tag{C8}$$

where $\gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) = 1 - \gamma_1\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right)$. The EM algorithm to estimate equation (C1) then simply amounts to repeating the following two steps for $r = 1, 2, \dots$ until convergence is reached, starting with initial values $\alpha^{(0)}$ and $\boldsymbol{\pi}^{(0)}$:

(i) Set

$$\alpha^{(r+1)} = \frac{1}{T} \sum_{t=1}^{T} \gamma_1\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right). \tag{C9}$$

(ii) Set

$$\boldsymbol{\pi}^{(r+1)} = \left(\frac{\sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{X_t \in [b_{k-1}, b_k]}}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{X_t \in [b_{l-1}, b_l]}}, \dots, \frac{\sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{X_t \in [b_{k-1}, b_k]}}{\sum_{l=1}^{K} \sum_{t=1}^{T} \gamma_2\left(x_t; \boldsymbol{\pi}^{(r)}, \alpha^{(r)}\right) \cdot \mathbb{1}_{X_t \in [b_{l-1}, b_l]}} \right). \tag{C10}$$

C2 For unsupervised classification

For full generality, we assume the data $\boldsymbol{X}_{1:T} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_T)$ is an independent and identically distributed sample from a mixture of K multivariate normal distributions

$$F = \sum_{k=1}^{K} \alpha_k \cdot \mathcal{N}_d(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{C11}$$

where $\alpha_1, \dots, \alpha_K \in (0, 1)$ with $\sum_{k=1}^{K} \alpha_k = 1$. We define the independent and identically distributed latent variables $Z_{1:T} = (Z_1, \dots, Z_T)$ such that $Z_t \sim \text{Categorical}(K; \boldsymbol{\alpha})$ (i.e. each Z_t is a discrete $\{1, \dots, K\}$-valued random variable with $\mathbb{P}(X_t = k) = \alpha_k$) and $\mathbf{X}_t \mid (Z_t = k) \sim \mathcal{N}_d(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. Writing $\boldsymbol{\eta} = (\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K, \boldsymbol{\alpha})$, the complete-data log-likelihood is then

$$\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{x}, Z_{1:T}) = \log\left[\prod_{t=1}^{T} \prod_{k=1}^{K} \left(\frac{\alpha_k}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}_k|}} \cdot \exp\left(-\tfrac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_k)\right)\right)^{\mathbb{1}_{Z_t=k}}\right] \tag{C12}$$

$$= \sum_{t=1}^{T} \sum_{k=1}^{K} \mathbb{1}_{Z_t=k} \cdot \left[\log(\alpha_k) - \tfrac{1}{2}\left(d\log(2\pi) + \log(|\boldsymbol{\Sigma}_k|) + (\boldsymbol{x}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_k)\right)\right]. \tag{C13}$$

The E-step requires the computation of $\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{X}_{1:T}, Z_{1:T}) \mid \boldsymbol{X}_{1:T} = \boldsymbol{x}_{1:T}]$, which this time requires $\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[\mathbb{1}_{Z_t=k} \mid \boldsymbol{X}_{1:T} = \boldsymbol{x}_{1:T}] = \mathbb{P}_{\boldsymbol{\eta}^{(r)}}(Z_t = k \mid \boldsymbol{X}_t = \boldsymbol{x}_t)$ to be computed. Again, Bayes' rule and the law of total probability yield

$$\mathbb{P}_{\boldsymbol{\eta}^{(r)}}(Z_t = k \mid \boldsymbol{X}_t = \boldsymbol{x}_t) = \frac{\alpha_k^{(r)} \cdot \phi_d\left(\boldsymbol{x}_t; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}\right)}{\sum_{l=1}^{K} \alpha_l^{(r)} \cdot \phi_d\left(\boldsymbol{x}_t; \boldsymbol{\mu}_l^{(r)}, \boldsymbol{\Sigma}_l^{(r)}\right)} =: \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right), \tag{C14}$$

where $\phi_d(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the $\mathcal{N}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density function. The M-step thus requires the maximization of

$$\mathbb{E}_{\boldsymbol{\eta}^{(r)}}[\ell_{\text{com}}(\boldsymbol{\eta} \mid \boldsymbol{X}_{1:T}, Z_{1:T}) \mid \boldsymbol{X}_{1:T} = \boldsymbol{x}_{1:T}] = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right) \cdot \left[\log(\alpha_k) - \frac{1}{2}\left(d\log(2\pi) + \log(|\boldsymbol{\Sigma}_k|) + (\boldsymbol{x}_t - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_k)\right)\right] \tag{C15}$$

with respect to each $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and $\boldsymbol{\alpha}$. It is straightforward to show that the maximizing value of α_k is

$$\hat{\alpha}_k = \frac{1}{T} \sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right). \tag{C16}$$

The remaining parameters are most easily optimized using matrix calculus (we omit details but see e.g. Muirhead 2009), which yield the optima

$$\hat{\boldsymbol{\mu}}_k = \frac{1}{\sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right)} \sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right) \boldsymbol{x}_t \tag{C17}$$

and

$$\hat{\boldsymbol{\Sigma}}_k = \frac{1}{\sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right)} \sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right) (\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_k)(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}}_k)^\top. \tag{C18}$$

The EM algorithm to estimate (C1) then simply amounts to repeating the following two steps for $r = 1, 2, \dots$ until convergence is reached, starting with initial values $\boldsymbol{\alpha}^{(0)}$ and $\boldsymbol{\mu}_1^{(0)}, \dots, \boldsymbol{\mu}_K^{(0)}, \boldsymbol{\Sigma}_1^{(0)}, \dots, \boldsymbol{\Sigma}_K^{(0)}$:

(i) Set

$$\boldsymbol{\alpha}^{(r+1)} = \left(\frac{1}{T} \sum_{t=1}^{T} \gamma_1\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right), \dots, \frac{1}{T} \sum_{t=1}^{T} \gamma_K\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right)\right). \tag{C19}$$

(ii) Set

$$\boldsymbol{\mu}_k^{(r+1)} = \frac{1}{\sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right)} \sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right) \boldsymbol{x}_t, \quad k = 1, \dots, K. \tag{C20}$$

(iii) Set

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{1}{\sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right)} \sum_{t=1}^{T} \gamma_k\left(\boldsymbol{x}_t; \boldsymbol{\eta}^{(r)}\right) \left(\boldsymbol{x}_t - \boldsymbol{\mu}_k^{(r+1)}\right) \left(\boldsymbol{x}_t - \boldsymbol{\mu}_k^{(r+1)}\right)^\top, \quad k = 1, \dots, K. \tag{C21}$$

The initial values for mixture models such as equation (C11) are typically obtained using the k-means algorithm. In the present case, one can run this algorithm (available in any statistical software package) on $\boldsymbol{X}_{1:T}$ with the number of centres specified as K, which partitions the data into K distinct subsets; for each mixture component $k \in \{1, \dots, K\}$, $\boldsymbol{\mu}_k^{(0)}$ and $\boldsymbol{\Sigma}_k^{(0)}$ are respectively set to the sample mean and covariance matrix from subset k, while $\alpha_k^{(0)}$ is set to the proportion of $\boldsymbol{X}_{1:T}$ that comprises subset k.

Table D1. Maximum-likelihood estimates for Model 1 fit using ObsID 01885 with $w \in \{25, 50, 75, 100\}$.

	$w = 25$	$w = 50$	$w = 75$	$w = 100$
ϕ_1	0.9874	0.9755	0.9636	0.9563
σ_1	0.0794	0.1161	0.1383	0.1576
β_1	0.1769	0.1787	0.1861	0.1823
β_2	0.0726	0.0733	0.07636	0.0748

Table D2. Maximum-likelihood estimates for Model 2 fit using ObsID 01885 with $w \in \{25, 50, 75, 100\}$.

	$w = 25$	$w = 50$	$w = 75$	$w = 100$
ϕ_1	0.9883	0.9773	0.9672	0.9591
σ_1	0.0667	0.0961	0.1147	0.1309
σ_2	0.1069	0.1539	0.1843	0.2099
β_1	0.1872	0.1864	0.1921	0.1869
β_2	0.0597	0.0593	0.0622	0.0596

Table D3. Maximum-likelihood estimates for Model 3 fit using ObsID 01885 with $w \in \{25, 50, 75, 100\}$.

	$w = 25$	$w = 50$	$w = 75$	$w = 100$
ϕ_1	0.9884	0.9768	0.9693	0.9612
ϕ_2	0.9878	0.9744	0.9647	0.9549
σ_1	0.0711	0.0987	0.1143	0.1305
σ_2	0.1064	0.1568	0.1905	0.2152
β_1	0.1865	0.1860	0.1877	0.1836
β_2	0.0596	0.0591	0.0594	0.0580
ρ	1.0000	1.0000	1.0000	1.0000

Note that when a mixture model involves all component distributions within the same parametric family (such as equation C11, but not equation C1), any permutation of the component 'labels' $1, \ldots, K$ produces the same value of the ordinary log-likelihood function, in the sense that

$$\ell(\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K \mid \boldsymbol{X}_{1:T}) = \ell(\boldsymbol{\lambda}_{\sigma(1)}, \ldots, \boldsymbol{\lambda}_{\sigma(K)} \mid \boldsymbol{X}_{1:T}), \tag{C22}$$

where $\boldsymbol{\lambda}_k$ is the set of parameters associated with the kth component distribution (including the mixing parameter α_k) and σ is any permutation of $(1, \ldots, K)$. This is an example of a phenomenon known as unidentifiability, which results in K modes in the log-likelihood surface; the value of the log-likelihood at each such mode is the same, and so the EM algorithm can converge to any one of them. Thus, from a computational perspective, one cannot a priori associate any particular physical state (such as quiescence) to a specific component distribution of equation (C11).

When the $\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_K$ can be ordered in some way, any particular ordering of the 'labels' can be imposed on the likelihood function; for example, if equation (C11) comprises of K univariate normal distributions and one desires the component distributions to be ordered increasingly with respect to their means, then one can set the log-likelihood to $-\infty$ whenever $\mu_1 < \mu_2 < \cdots < \mu_K$ fails to hold. Alternatively, the EM algorithm can sometimes be coaxed towards a particular labelling by judiciously choosing initial values. When K is small, however, one may assign meaning to the components settled on by the algorithm following estimation. If one desires a specific ordering of the components, the labels of the estimated parameters can simply be permuted.

APPENDIX D: ADDITIONAL RESULTS FOR EV LAC

D1 Model estimates for varying time bins

In this section, we provide maximum-likelihood estimates from the first stage of the model estimation procedure (prior to the bootstrap-based de-biasing procedure) for each of the three models described in Section 4.3 fit to ObsID 01885, as the time bin w (in seconds) varies among $\{25, 50, 75, 100\}$. The estimates are given in Tables D1, D2, and D3, respectively and do not vary materially with w. Our experiments show that the de-biased estimates are similarly stable.

We do observe that in general, the σ_1 and σ_2 increase steadily as w increases. This behaviour is expected, because these parameters control the step size (per unit time) of the underlying Markov chain. Heuristically, suppose that $X_{1:T}$ is the underlying soft-band process associated with the original time bin w, and $X'_{1:T'}$ is that associated with a larger time bin w', where $T' < T$. Then the resolution of $X'_{1:T'}$ is lower than that of $X_{1:T}$, and so within each time bin of length w', $X_{1:T}$ takes several independent steps (say s of them, where $s > 1$) while $X'_{1:T'}$ takes a single step; that is, $X_{t_1:t_s}$ occur at the same time as X'_t, and the error of the latter is approximately the sum of the errors of each component

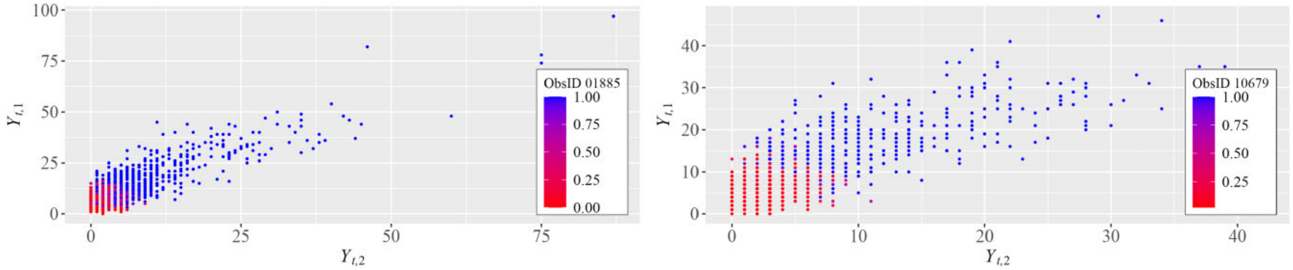

Figure D1. Posterior flaring state probabilities computed via equation (28) for ObsID 01885 and via equation (30) for ObsID 10679. The probabilities are plotted as a function of the observed soft-band counts $Y_{1,1}, \ldots, Y_{T,1}$ and the observed hard-band counts $Y_{1,2}, \ldots, Y_{T,2}$ for ObsID 01885 (left) and ObsID 10679 (right). Colour represents the posterior probability of the flaring state. It would not be possible for a flux threshold to reproduce these probabilities.

of $X_{t_1:t_s}$. If σ_1 and σ_1' are the parameters associated with $X_{1:T}$ and $X'_{1:T'}$, respectively, then $\sigma_1' \approx \sqrt{s}\sigma_1 > \sigma_1$. (This argument makes several simplifying assumptions, but can be made rigorous.)

D2 Scatterplots of hard and soft counts

Scatterplots of the hard and soft counts coloured according to their posterior probabilities of being associated with the flaring state of EV Lac (ObsID 01885 and ObsID 10679) appear in Fig. D1. The scatterplots confirm that these probabilities could not be obtained with a threshold on the observed counts.

This paper has been typeset from a TEX/LATEX file prepared by the author.