Scalable Bayesian Inference for Probabilistic Spectrotemporal Models of Type Ia Supernovae

Ana Sofía Uzsoy

Department of Engineering
University of Cambridge

This dissertation is submitted for the degree of
Master of Philosophy

Churchill College August 2022

Dedicated to all of the amazing people from all walks of life that I have met during my year in Cambridge.

Declaration

I, Ana Sofía Uzsoy of Churchill College, being a candidate for the MPhil in Machine Learning and Machine Intelligence, hereby declare that this report and the work described in it are my own work, unaided except as may be specified below, and that the report does not contain material that has already been used to any substantial extent for a comparable purpose.

All code used in this thesis is my own original work implementing the model described in Mandel et al. (2022) and Thorp et al. (2021), with the exception of basic cubic spline utilities written by S. Thorp. The MCMC chains for the Foundation dataset used for comparison in Chapter 5 were generated by Thorp et al. (2021). This work utilized Python packages such as `Pyro` (Bingham et al., 2018), `PyTorch` (Paszke et al., 2019), `astropy` (Astropy Collaboration et al., 2013), `corner` (Foreman-Mackey, 2016), `extinction`, and standard Python packages. All code used in this work is available on GitHub at `https://github.com/asmuzsoy/variational_bayesn`.

The dissertation contains 11,371 words, including captions.

Ana Sofía Uzsoy
August 2022

Acknowledgements

I would like to thank my supervisors Dr. Kaisey Mandel and Stephen Thorp for their guidance and feedback. I would also like to acknowledge the Astronomy and Engineering departments and especially the course administrators of the MLMI MPhil for their support throughout the year. I am also extremely grateful to the Churchill Scholarship for providing my funding to complete this course.

Abstract

Type Ia supernovae serve as “standard candles" whose observed light curves (time series of brightness across wavelengths) allow for accurate calculation of their distances from Earth. This provides rare insight into the rate of the expansion of the universe and other fundamental cosmological quantities. As technology advances, vast quantities of data are generated from astronomical observations, and principled, scalable modeling approaches are required for effective analysis. BayeSN is a generative, hierarchical Bayesian model that fits the light curves of Type Ia supernovae, allowing for inference of their distance and other parameters. The current implementation of BayeSN is written in Stan and uses Hamiltonian Monte Carlo to sample the posterior distribution, which is limited by computation time. We present an implementation of BayeSN that uses Variational Inference (VI) to fit Type Ia supernova light curves. This model is implemented in Pyro and utilizes Stochastic Variational Inference to optimize a full-rank multivariate Gaussian approximation of the posterior distribution. We present fit results on simulated and real supernova datasets and discuss the trade-off between computational efficiency and model accuracy. We compare the performance of the VI model to previous MCMC approximations and discuss the trends and biases of the VI model. We also explore how the model can be reparameterized using different variable transformations to improve performance. Overall, VI is a promising method for efficiently fitting Type Ia supernova light curves, and machine learning techniques should continue to be utilized to accelerate physically motivated models to advance the field of precision cosmology.

Table of contents

List of figures **viii**

Nomenclature **x**

1 Introduction **1**

2 Background **3**

2.1 Primer on astronomical data and cosmology 3

2.1.1 Magnitudes and distances 3

2.1.2 Redshifts and cosmology 4

2.1.3 Dust 5

2.2 Type Ia supernovae and precision cosmology 7

2.3 Computational techniques in astronomy 8

2.4 Overview of variational inference 9

2.5 Introduction to Pyro 11

3 Overview of model & VI implementation **13**

3.1 What is BayeSN? 13

3.2 The BayeSN generative model 14

3.3 Generative model implementation details 15

3.4 Description of single-supernova Pyro model 17

4 VI model results on simulated type Ia supernova light curves **21**

4.1 Method to simulate light curves 21

4.2 Performance on simulated data 22

4.2.1 Single simulated supernova 22

4.2.2 Simulated supernova population 23

5 Results on Foundation Dataset & Comparison with MCMC 27

5.1 Overview of dataset 27

5.2 Methods 27

5.3 Comparison of results to expected Hubble relation 28

5.4 Residuals analysis 29

6 Investigating & mitigating bias in dust extinction 34

6.1 Motivation 34

6.2 Softplus transformation 34

6.2.1 Theoretical overview 34

6.2.2 Implementation 35

6.3 Asymmetric Laplace Distribution 38

6.3.1 Theoretical overview 38

6.3.2 Implementation 40

7 Conclusion 44

References 48

Appendix A Derivation of Laplace Approximation 52

Appendix B Distance modulus relation 53

Appendix C Code 54

List of figures

2.1 Bandpass filters over SED throughput 4
2.2 Dust extinction over wavelength bands 6

3.1 PGM of BayeSN from Mandel et al. (2022) 14
3.2 BayeSN schematic from Mandel et al. (2022) 16
3.3 PGM of BayeSN-VI Model 19

4.1 Simulated light curve fit 23
4.2 Comparison of fits for simulated light curve 24
4.3 Fit vs. true parameters for simulated supernovae 25
4.4 Triangle plot of simulation residuals 26

5.1 Hubble diagram with Foundation dataset 28
5.2 Hubble residuals for Foundation dataset 29
5.3 Residual plot for Foundation A_V 30
5.4 Residual plot for Foundation μ 31
5.5 Residual plot for Foundation θ 32
5.6 Triangle plot for residuals for Foundation dataset 33

6.1 PDF of softplus-transformed parameter x 36
6.2 Softplus-tranformed distribution for simulated supernova with low A_V . . . 37
6.3 Fit vs. true x (transformed value) 38
6.4 Simulated population fit using softplus transformation 39
6.5 MCMC A_V vs. VI Softplus A_V 39
6.6 VI Softplus A_V Residuals 39
6.7 PDF of Asymmetric Laplace distribution 41
6.8 Asymmetric Laplace distribution for simulated supernova with low A_V . . . 42
6.9 Simulated population fit using asymmetric Laplace prior 43
6.10 MCMC A_V vs. VI Asymmetric Laplace A_V 43

6.11 Asymmetric Laplace A_V Residuals 43

7.1 PGM of population-level BayeSN model 46

Nomenclature

Acronyms

ELBO	Evidence-Based Lower Bound
FPC	Functional Principal Components
MCMC	Markov Chain Monte Carlo
MLE	Maximum Likelihood Estimate
PDF	Probability Density Function
PPL	Probabilistic Programming Language
SED	Spectral Energy Distribution
SN/SNe	Supernova/Supernovae
SVI	Stochastic Variational Inference
VAE	Variational Auto-Encoder
VI	Variational Inference

Symbols

μ	the distance modulus, a parameter in the VI model
ν	the residual covariance parameter in the VI model
θ	the coefficient for the first FPC, a parameter in the VI model
A_V	the dust extinction coefficient, a parameter in the VI model
g,r,i,z	names of wavelength ranges (bandpass filters)

H_0 the Hubble constant, the rate of expansion of the universe

z redshift

Astronomical terminology

absolute magnitude the brightness of an object if it were a distance of 10 pc away

apparent magnitude the brightness of an object as it is observed

distance modulus the difference of apparent and absolute magnitude, denotes distance

extinction The dimming of observed light from an astronomical object due to dust

Hubble constant the present-day rate of expansion of the universe

light curve the flux over time of a supernova, usually measured within wavelength bands

pc parsec, a unit of distance equal to 3.26 light years

redshift Recessional velocity of an astronomical object, in units of the speed of light

Chapter 1

Introduction

Supernovae are stellar explosions so bright they can be detected from the Earth across great cosmic distances. Type Ia supernovae are unique in that their absolute brightness can be standardized; that is, their observed brightness profile correlates directly with their distance from the Earth. Accurately modeling the time-series of brightness at different wavelengths for a Type Ia supernova population enables astronomers to fit their distances, which can be used to estimate various cosmological quantities, such as the rate of expansion of the universe. Future astronomical surveys plan to observe millions of supernovae, creating a distinct need for precise, scalable supernova light curve modeling techniques.

BayeSN is a forward generative, hierarchical Bayesian model for the time-dependent spectral energy distribution of Type Ia SNe (Mandel et al., 2022; Thorp et al., 2021). It is currently implemented in Stan, and uses Hamiltonian Monte Carlo to sample the posterior distribution. While this approach provides an accurate estimate of latent variables, it is quite computationally expensive and is thus limited in the scale of the analyses it can complete.

This work replaces the Monte Carlo sampling of the posterior with variational inference (VI) to enable accelerated estimation of the posterior distribution while still retaining the robustness of the hierarchical Bayesian model. We find generally good agreement with true parameters in simulation-based and with MCMC fits for real supernova. We find slight biases in the estimation of dust extinction and explore different modifications to the model that help mitigate this effect. These contributions provide a building block towards scalable hierarchical inference for faster, more accurate estimates of cosmological parameters.

This thesis provides a description of the VI model and an analysis of its performance on various real and simulated Type Ia supernova datasets. Chapter 2 provides necessary background information, including a primer on cosmology, an overview of the role of computational techniques in astronomy, and an explanation of variational inference (VI) and how it is implemented in Pyro. Chapter 3 provides a detailed walkthrough of the BayeSN generative hierarchical Bayesian model and how the VI model is implemented in practice. Chapter 4 describes how supernova light curves are simulated, and compares the model performance on these light curves to the known true parameter values. Chapter 5 describes the model performance on the Foundation dataset, which consists of 157 real observed supernova light curves, and compares the results of the VI model to previously determined MCMC results. Chapter 6 provides a deeper look into potential biases in the inference of the effects of cosmic dust and explores different modifications to the model and their effectiveness in mitigating this bias. Finally, Chapter 7 presents overall conclusions as well as a discussion of potential future work.

Chapter 2

Background

2.1 Primer on astronomical data and cosmology

2.1.1 Magnitudes and distances

The logarithmic scale along which the brightness of an astronomical object are measured is known as magnitude. The apparent magnitude denotes the brightness of an object as seen from Earth, while the absolute magnitude represents the brightness that would be measured if the object was a distance of 10 pc away from Earth. In this way, the apparent magnitude represents the observed brightness of an object, while the absolute magnitude represents the intrinsic brightness (Hogg, 2022).

The absolute magnitude M is related to the distance modulus μ and apparent magnitude m as follows:

$$\mu = m - M \tag{2.1}$$

where μ can be used to determine the distance d to the object:

$$\mu = 5\log_{10}(\frac{d}{10\text{pc}}) \tag{2.2}$$

An astronomical object's Spectral Energy Distribution (SED) denotes the distribution of its light (or energy) over different wavelengths. Brightness can measured within “bandpass filters" denoted by single letters (u, b, g, r, etc) that span certain wavelength ranges. Figure 2.1 shows a Type Ia supernova SED and the transmission functions for the $g, r, i,$ and z bands (which are utilized later in this work). To calculate the magnitude of an object within a given bandpass filter, the SED must be integrated with the corresponding transmission function in wavelength space (Hogg, 2022).

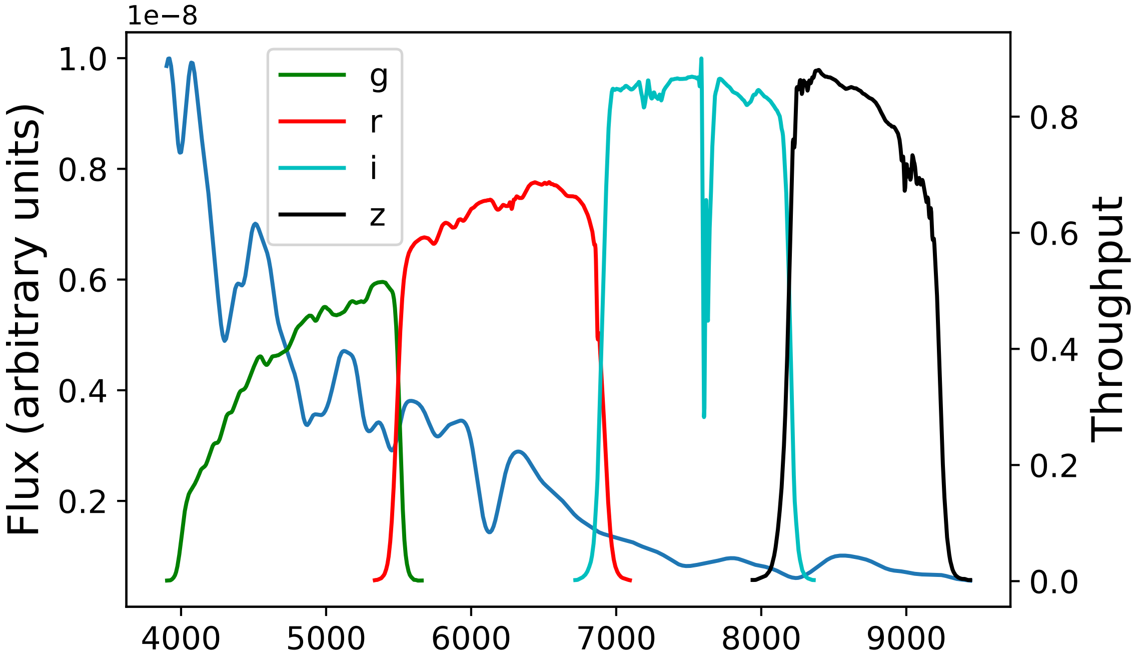

Fig. 2.1 Type Ia supernova SED at $t = 0$ (blue), determined as an average of many supernovae via the Hsiao template (Hsiao et al., 2007). The g, r, i, and z transmission functions are overlaid on top. An object's magnitude in each filter is determined by integrating the SED under the transmission function (Hogg, 2022).

The magnitude is related to the observed flux f and flux density F from the object SED as follows:

$$m = -2.5\log_{10}(f) + Z \tag{2.3}$$

$$= -2.5\log_{10}\int F(\lambda)\mathbb{B}(\lambda)\lambda \mathrm{d}\lambda \tag{2.4}$$

where Z = 27.5 by convention and $\mathbb{B}(\lambda)$ denotes the transmission function for particular bandpass filter, as seen in Figure 2.1 (for more details see Mandel et al. (2022), Section 2.1).

2.1.2 Redshifts and cosmology

The idea that objects in the universe were actively moving away from us, and were receding faster the further they were, was first proposed in the 1920s (Lemaître, 1927; Hubble, 1929). Since then, the redshift z of an object, which denotes its recessional velocity (i.e. how fast it is moving away from us) has been observed and recorded across virtually all kinds of astronomical objects (Lemaître, 1927; Hubble, 1929). This movement is quantified with the

Hubble constant H_0 serves as the constant of proportionality between the redshift and the distance to the object (Hubble, 1929):

$$cz \approx H_0 d \tag{2.5}$$

The object's distance modulus μ can be determined from its observed redshift with the following relation:

$$\begin{aligned} \mu &= 25 + 5\log_{10}\left[\frac{c}{H_0}\tilde{d}_L(z_s;\Omega_M,\Omega_\Lambda,w)\right]\mathrm{Mpc}^{-1} && (2.6)\\ &\approx 25 + 5\log_{10}\frac{cz}{H_0 \times \mathrm{Mpc}} && (2.7)\end{aligned}$$

where c is the speed of light (3×10^5km/s) (see e.g. Avelino et al., 2019, eq. 2). Equation 2.7 is a linear approximation of Equation 2.6 at low redshifts (Hogg, 1999). A more detailed explanation of Equation 2.6 is available in Appendix 2.

2.1.3 Dust

Interstellar dust consists of small grains of silicate materials, around $10^{-6} - 10^{-10}$ meters wide, floating throughout the universe (Draine, 2003). In any astronomical observation, one must take into account the effects of dust, which can affect the measured brightness of the object being observed in different wavelengths. The dust that most affects observations is interstellar dust along the line of sight to the object our own Milky Way galaxy and the object's host galaxy (Draine, 2003). The dimming effect of dust is known as “dust extinction".

Dust extinction occurs everywhere, but is usually stronger in wavelength regions closer to the blue end of the visual spectrum. This effect is also known as “reddening" due to this differential dimming effect causing spectra to be biased towards red wavelengths.

The extinction coefficient A_V denotes the dimming of the observed magnitude in the V wavelength band, which corresponds to visible light (~551 nm). The ratio

$$R_V = \frac{A_V}{(A_B - A_V)} \tag{2.8}$$

where B is another wavelength band (~445 nm), denotes the overall rate of extinction in visible wavelengths (Draine, 2003).

We use dust extinction laws to describe the wavelength-dependent dimming effects of dust (Cardelli et al., 1989). Figure 2.2 shows the dust extinction curves derived from the Fitzpatrick (1999) law for different values of R_V, as well as the different wavelength regions for bands g, r, i, and z. Note the different extinction behavior in the different bands- generally, observations from as many bands as possible are used to constrain the value of R_V. The value of R_V corresponds roughly to the slope of the curve, with low values having a much higher difference in slope between bands (as expected, by maximizing the denominator of Equation 2.8) and higher values having more consistent slope along all wavelengths. The cyan line denotes $R_V = 2.61$, a value inferred by Thorp et al. (2021) for the SN Ia sample used in probabilistic models later in this work.

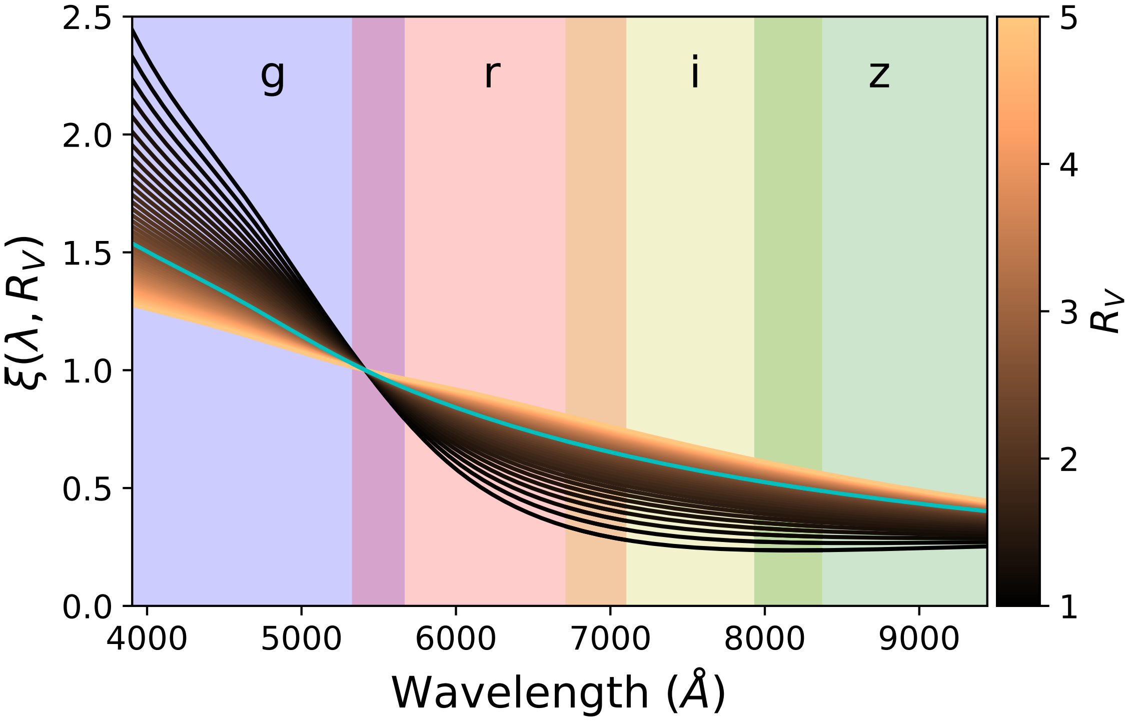

Fig. 2.2 The Fitzpatrick (1999) dust extinction law, calculated for values of R_V ranging from 1 to 5, with the wavelength band g, r, i, and z regions colored in blue, pink, yellow, and green respectively. The cyan line indicates R_V = 2.61, the best fitting value inferred by Thorp et al. (2021) on the Foundation dataset, used later in this work.

Note that while we have a good understanding of Milky Way dust (Schlafly & Finkbeiner, 2011; Schlafly et al., 2016; Schlegel et al., 1998), and dust laws in general, we do not have detailed knowledge of the dust of distant galaxies, thus needing to model the dust extinction probabilistically.

2.2 Type Ia supernovae and precision cosmology

When stars reach the end of their lifetimes, they become supernovae and explode, creating heavy elements and emitting bright light across many wavelengths (Jha et al., 2019). Using ground- and space-based telescopes, their optical properties (i.e. magnitudes in different wavelength bands) and spectral properties (i.e. signatures of different elements in their chemical composition) can be observed. Supernovae are classified according to these characteristics, with SNe Ia having a unique secondary peak in the infrared and spectral signatures of calcium, magnesium, silicon, and sulfur (Jha et al., 2019).

Type Ia SNe have the intriguing characteristic that they are “standardizeable candles", meaning that their absolute magnitudes are uniform. To determine the standardized absolute magnitudes M, different corrections must be applied to these observed luminosities in each filter, as well as overall corrections for host galaxy properties (Phillips, 1993; Jha et al., 2019). Using standardized magnitudes, the distance to a supernova can be determined from its apparent magnitudes m (as seen in Equations 1.1 and 1.2), with closer objects appearing brighter and farther objects appearing dimmer.

While the precision of SNe Ia as distance indicators has many important implications, one of the most prominent is for the expansion of the universe. Riess et al. (1998) and Perlmutter et al. (1999) used SNe Ia to precisely show that this expansion rate itself evolved over time and that the universe is expanding outwards at an accelerating rate. Freedman et al. (2001), was able to estimate the Hubble constant H_0, which denotes present-day the rate of the universe’s expansion.

Since this landmark work, H_0 has been estimated hundreds of times using data from many different astronomical objects, including supernovae, gravitational waves, and the Cosmic Microwave Background, with special attention being paid to the measurement errors. There is still a lack of general consensus on a value and uncertainty range for H_0, known as the “Hubble tension" (Knox & Millea, 2020; Freedman, 2021). The SH0ES project uses Type Ia supernovae to measure $H_0 = 73.04 \pm 1.04$ km/s/Mpc, while the Planck Collaboration uses the Cosmic Microwave Background and finds $H_0 = 67.4 \pm 0.5$ (Riess et al., 2022; Planck Collaboration et al., 2020). This tension is one of the most prominent unsolved problems in modern cosmology.

Significant work has been done attempting to dissolve this tension. Since much of the original analysis was done using optical wavelengths, Dhawan et al. (2018) and Burns et al. (2018) use near-infrared wavelengths to calculate H_0. Feeney et al. (2018) uses a hierarchical Bayesian model to infer H_0 using the same data from the Riess et al. (1998) and Perlmutter et al. (1999). Still, there is no firm consensus on the Hubble constant; thus, improved methods and continued analysis are necessary.

2.3 Computational techniques in astronomy

As increasingly elaborate ground and space-based telescopes are developed and deployed, the amount of astronomical data grows accordingly. Current and future surveys, such as the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST) and Nancy Grace Roman Space Telescope, will observe objects at further distances and higher redshifts than ever before in hopes of characterizing dark energy, which is thought to be the driving force behind the expansion of the universe. But, the methods used to analyze astronomical observations must grow simultaneously with technological advancements, and the need for robust, scalable modeling will only increase with the amount of data (Breivik et al., 2022; Ivezić et al., 2019; Rose et al., 2021).

The LSST, which began surveying the sky in 2022, is expected to make over 32 trillion observations in over 20 billion galaxies and stars, many over time (creating time-series data) (The LSST Dark Energy Science Collaboration et al., 2018). Breivik et al. (2022) outline the data analysis requirements for the LSST, including cross-matching new observed sources with existing catalogued observations, creating selection functions to determine population statistics, and analysis of time-series data and images. New computational techniques such as machine learning can help astronomers explore this influx of data, especially when it comes to analyzing population statistics of astronomical objects (Dvorkin et al., 2022).

There are several examples of applying machine learning and related techniques to supernova data. Huber et al. (2022) use neural networks and random forests to measure the time delay of strongly lensed type Ia supernovae. Kim et al. (2013) and Vincenzi et al. (2019) apply Gaussian processes to observed supernova data over time.

Boone (2021) uses a variational autoencoder to fit Type Ia SNe light curves with training solely on photometric observations. Their latent representation of each supernova can be mapped back to physical characteristics, with the distance to each supernova being calculated as a linear combination of the three latent variables (Boone, 2021). This model, implemented in PyTorch, was able to replicate the results of more traditional supernova light curve analyses while significantly reducing the computation time, which demonstrated the potential of modern machine learning methods for performing scalable inference in this field (Boone, 2021).

In many cases, astronomers use Markov Chain Monte Carlo (MCMC) to sample a complicated posterior. The Python package `emcee` in particular has become ubuquitous among astronomical literature in recent years (Foreman-Mackey et al., 2013). MCMC provides the advantage that the form of the posterior does not have to be known, and that the posterior can be well approximated with sampling. However, MCMC can often be computationally inefficient and require large amounts of computation time and resources. Its runtime is also directly correlated with the number of samples it has to produce.

Variational inference (VI) presents an alternate, more computationally efficient way to approximate a posterior. Regier et al. (2018) find that VI is 1,000 times faster than MCMC for cataloguing optical images from telescopes. A VI-based method for distinguishing individual stars in blended images performed as much as 100,000 times faster than a competing method using MCMC (Liu et al., 2021). In using VI for precision cosmology, Rizzato & Sellentin (2022) find that it uses only 0.6% of the numerical cost compared with MCMC.

Using VI to increase scalability not only saves time and computational resources, but also provides a valuable basis for comparison for existing methods. Additionally, using VI allows for use of more elaborate models that are currently limited by computational time, providing additional insight into astrophysical systems. For the specific problem of Type Ia supernovae, VI will allow for faster analysis of more supernovae, and for the use of increasingly complex population and dust models.

2.4 Overview of variational inference

Variational inference (VI) is a unique machine learning approach that approximates a complex, often intractable posterior distribution with a simpler “surrogate" posterior. The method aims to optimize a given form of posterior distribution such that it minimizes the difference between the surrogate posterior and the true posterior (Jordan et al., 1999).

The similarity between two distributions can be evaluated with the Kullback-Leibler (KL) divergence, which measures the flow of information between two distributions and is as follows (Kullback & Leibler, 1951):

$$D_{KL}(p(x) \,||\, q(x)) = \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx \tag{2.9}$$

In Bayesian inference, for data x and model z, we use the likelihood $p(x|z)$, the prior $p(z)$ and the marginal likelihood $p(x)$ to determine the posterior distribution $p(z|x)$:

$$p(z|x) = \frac{p(z,x)}{p(x)} \propto p(x|z)p(z) \tag{2.10}$$

The goal of Bayesian inference is usually to sample the posterior distribution of model parameters conditional on observed data. In some cases, the posterior distribution $p(z|x)$ is too complicated or intractable for conventional inference methods. VI seeks to minimize the KL divergence between the surrogate posterior $q^*(z)$ and the original posterior $p(z|x)$

(Jordan et al., 1999; Wainwright et al., 2008):

$$q^*(z) = \arg\min_{q(z)\in\mathcal{Q}} D_{KL}(q(z) \mid\mid p(z|x)) \tag{2.11}$$

where $\mathcal{Q}$ denotes a chosen family of distributions (Blei et al., 2016). By seeking to minimize the KL divergence, VI turns an inference problem into an optimization problem.

We can rewrite the KL divergence as:

$$\begin{aligned} D_{KL}(q(z) \mid\mid p(z|x)) &= \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z|x)] && (2.12)\\ &= \mathbb{E}[\log q(z)] - \mathbb{E}[\log p(z,x)] + \log(p(x)) && (2.13)\\ &= \log(p(x)) - \text{ELBO(q)} && (2.14) \end{aligned}$$

where the Evidence-Based Lower Bound (ELBO) is characterized as:

$$\begin{aligned} \text{ELBO}(q(z)) &= \mathbb{E}[\log p(z,x)] - \mathbb{E}[\log q(z)] && (2.15)\\ &= \mathbb{E}[\log p(x|z)] - D_{KL}(q(z) \mid\mid p(z)) && (2.16) \end{aligned}$$

The ELBO is equivalent to the KL divergence up to a constant. Instead of minimizing the KL divergence, which can be difficult to calculate, VI maximizes the ELBO to find the approximate posterior $q^*(z)$ (Blei et al., 2016; Jordan et al., 1999; Wainwright et al., 2008). The left term in Equation 2.16 maximizes the likelihood of the data, while the term on the right minimizes the KL Divergence between the surrogate posterior and the prior on z, encouraging similarity (Blei et al., 2016).

Kingma & Welling (2014) introduce the method of Stochastic Variational Inference (SVI), which allows the surrogate posterior to be optimized using batched gradient descent. A naive Monte Carlo gradient estimator for the ELBO (or any function of the latent variable $f(z)$) would be:

$$\nabla\mathbb{E}_{q(z)}[f(z)] = \mathbb{E}_{q(z)}\left[f(z)\nabla_{q(z)}\log(q(z))\right] \approx \frac{1}{L}\sum_{l=1}^{L} f(z)\nabla_{q(z_l)}\log q(z_l) \tag{2.17}$$

where $z_l \sim q(z)$, the surrogate posterior. In practice, this estimator has been found to be highly variable and leads to poor results (Paisley et al., 2012). To circumvent this, a “reparametrization trick" is used to express the estimator in a different, less variable way. Instead of sampling a value z_i directly from the posterior $q(z)$, a new variable $\tilde{z}_i$ is created using the transformation $g(\varepsilon)$ with ε drawn from some distribution $p(\varepsilon)$ (Kingma & Welling, 2014).

A classic example of this type of reparametrization is the univariate Normal case, where $q(x) = \mathrm{N}(\mu, \sigma^2)$. Then:

$$\tilde{z}_i = \mu + \sigma \varepsilon_i, \text{ for } \varepsilon_i \sim \mathrm{N}(0,1) \tag{2.18}$$

With this reparameterization, new estimators of the ELBO can be constructed:

$$\widetilde{\mathrm{ELBO}(q(z))} = \left(\frac{1}{L} \sum_{l=1}^{L} \log p(x|\tilde{z}_l) \right) - D_{KL}(q(z) \mid\mid p(z)) \tag{2.19}$$

$$= \frac{1}{L} \sum_{l=1}^{L} \log p(x, \tilde{z}_l) - \log q(\tilde{z}_l) \tag{2.20}$$

where the estimator in Equation 1.16 is used if the KL-Divergence can be evaluated analytically (Kingma & Welling, 2014). Generally, the gradient is computed using random batches from the dataset to save time and memory, and parameters are fit using traditional gradient descent (Kingma & Welling, 2014).

VI has been widely used for many machine learning applications. One of the most notable is the variational auto-encoder (VAE), which uses variational inference to produce latent representations of input data (Kingma & Welling, 2014). In this way, the model can generate any amount of data based on the latent representation. A prominent application of VI and VAEs is image generation in conjunction with generative adversarial networks (GANs); Larsen et al. (2016) use a VAE-GAN to generate images of human faces that are more realistic than images generated just using VAEs or GANs separately by using the discriminator to train the auto-encoded images. VAEs have also been used to generate novel function-specific molecular structures (Sanchez-Lengeling & Aspuru-Guzik, 2018), model thermodynamic diffusion (Ho et al., 2020), and control agents in reinforcement learning tasks (Hafner et al., 2019).

2.5 Introduction to Pyro

Probabilistic programming languages (PPLs) allow for specialized and concise articulation of random variables drawn from probability distributions and their relationships to each other and to observed data. Spiegelhalter et al. (1995) developed Bayesian inference Using Gibbs Sampling, or BUGS, one of the first widely available PPLs. Today, there are a wide variety of PPLs, some of which utilize existing programming languages, such as Turing.jl (Ge et al., 2018), and Tensorflow Probability (Dillon et al., 2017), and some which implement their own, such as Stan (Stan Development Team, 2018).

Developed by researchers at UberAI, Pyro is a probabilistic programming language built on top of PyTorch (Paszke et al., 2019). Pyro supports stochastic variational inference (SVI) and uses PyTorch's automatic differentiation infrastructure to quickly and effectively optimize models. For SVI, a generative model has a secondary structural model, called a "guide", that serves as the surrogate posterior (Bingham et al., 2018). Pyro has auto-generating guides, that allow the user to specify the desired form of the approximate posterior, such as a Gaussian or Delta, distribution, and auto-populate the necessary parameters and arguments for the guide.

Pyro was chosen for use in this project over other PPLs due to its concise syntax and strong infrastructure for implementing SVI. Its model structure allows users to easily and specify conditional independence and express relationship between random variables. It has built-in methods for maximizing the ELBO using gradient descent and autodiff, and also for generating probabilistic graphical models (all PGMs in this work were made using Pyro's built-in model rendering method). Overall, Pyro's distinct support for SVI and helpful methods for development, debugging, and visualization made it the superior choice for this work.

Chapter 3

Overview of model & VI implementation

3.1 What is BayeSN?

BayeSN is a state-of-the-art hierarchical Bayesian model for the time-dependent spectral energy distributions (SEDs) of Type Ia supernovae (Mandel et al., 2022; Thorp et al., 2021). The current BayeSN model builds on years of previous work developing hierarchical Bayesian models for fitting Type Ia supernova light curves (Mandel et al., 2009, 2011), and can model the SED in time and wavelength space.

A key aspect of the design is the separation of the supernova's intrinsic properties (such as its distance and redshift) and the circumstances of its observation (such as the effects of its distance). Additionally, it leverages physical knowledge to express well-known properties of supernovae (such as dust extinction) while also using empirical, data-driven approaches to model more stochastic characteristics of the observed data (such as residual scatter).

It is known that a small number of characteristics, such as Nickel mass, can explain most of the intrinsic variance in Type Ia supernovae SEDs (Kasen et al., 2009). But, several smaller effects, such as metallicity or the explosion shape, cause additional variance in the brightness profile (Kasen et al., 2009). BayeSN's unique approach combines a functional probabilistic principal component analysis with residual dust extinction laws and additional stochastic residual terms to best model the intrinsic supernova SED (Bishop, 2016).

The observed light curve of a supernova is the result of a combination of several characteristics from both the individual supernova and the general population of supernovae. The hierarchical Bayesian structure of BayeSN allows for distinction between these individual and population-level parameters. Each supernova has a distinct shape, distance, and dust extinction coefficient, while the entire collection of supernovae share the principal components that capture common modes of SED variation and population-level distributions of supernova parameters.

Figure 3.1 shows a probabilistic graphical model of BayeSN (Mandel et al., 2022). The population-level parameters are along the outside of the graph, and include the assumed values of cosmological parameters (Ω_M, w, etc), the dust law describing extinction vs. wavelength, the intrinsic mean W_0 and first functional principal component W_1, and a dust extinction population distribution. Parameters inside the plate ($s = 1,...,N_{SN}$) are unique to each supernova and are drawn from the given population-level distributions, These include include the dust extinction coefficient A_V^s, the first principal component score θ_s, the distance to the supernova μ_s, and a time-and wavelength-varying residual realization η_s. These parameters are combined with the observed redshift z_s to generate a latent SED, which is then integrated through passbands to generated a light curve.

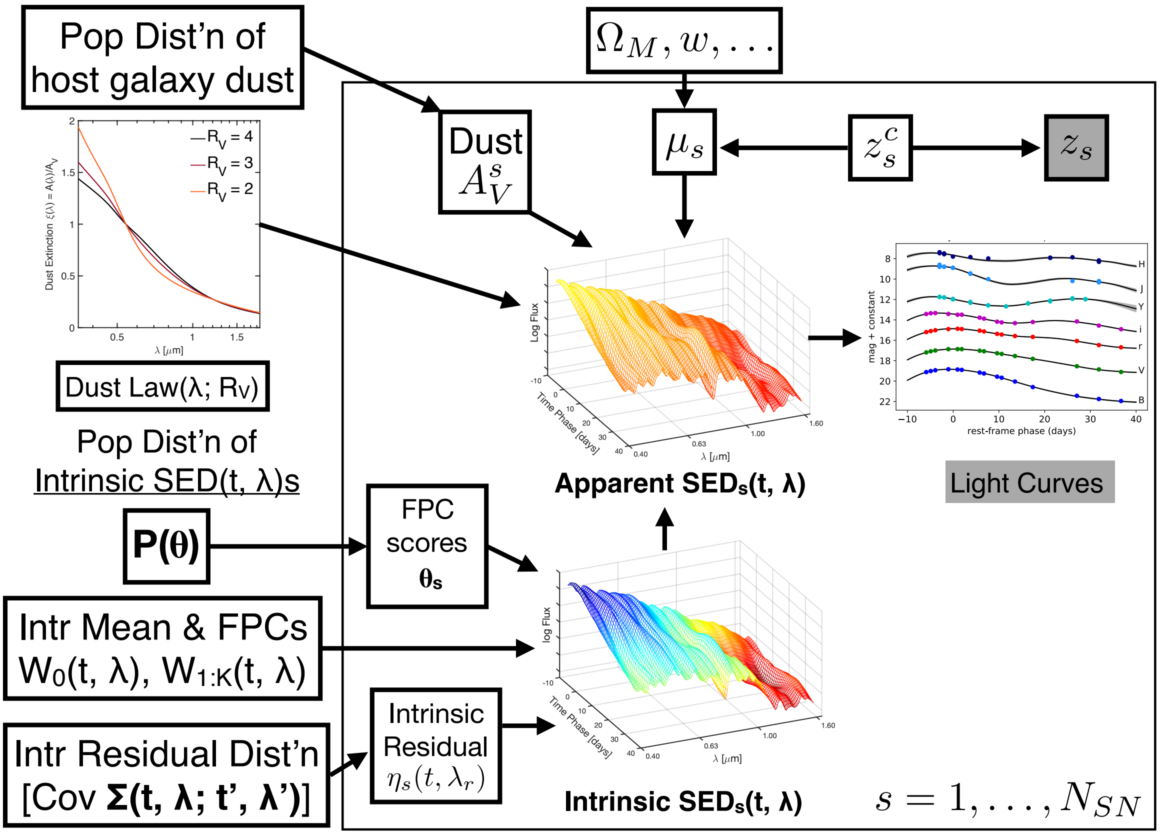

Fig. 3.1 A probabilistic graphical model of the parameters in BayeSN (Mandel et al., 2022). Arrows denote conditional dependency. Shaded boxes indicate observed parameters

3.2 The BayeSN generative model

The population and individual parameters combine as seen in Figure 3.2 to create the SED for each supernova (Mandel et al., 2022). The mean intrinsic SED is combined with the first principal component, weighted by the θ parameter. The dust extinction law is applied with the A_V parameter as a coefficient, and then a residual surface is fit to describe intrinsic SED perturbations beyond those captured by the functional principal components. The SED surface is then combined with the distance, redshift, and observation time points and integrated in different wavelength regions to create the light curves in different bandpass filters (Mandel et al., 2009, 2011, 2022).

As illustrated in Figure 3.2, the SED of a single supernova S_s can be written as a linear combination in log space (Mandel et al., 2022):

$$\begin{aligned} -2.5\log_{10}[S_s(t,\lambda_r)/S_0(t,\lambda_r)] = M_0 + W_0(t,\lambda_r) + \delta M_s + \left[\sum_{k=1}^{K} \theta_k^s W_k(t,\lambda_r)\right] \\ + \varepsilon_s(t,\lambda_r) + A_V^s \xi(\lambda_r;R_V) \end{aligned} \tag{3.1}$$

where S_0 is the Hsiao spectral template, which averages many supernovae into one SED (Hsiao et al., 2007; Hsiao, 2009), $M_0 = 19.5$, W_k is the k^{th} principal component of the SED with coefficient θ_k, $A_V^s\xi(\lambda_r;R_V)$ represents the effects of dust, δM_s includes uncertainty in μ_s, and $\varepsilon_s(t,\lambda_r)$ denotes the residual SED function.

Equation 6 from Mandel et al. (2022) shows how the generated flux f in wavelength band i is calculated from the SED:

$$\begin{aligned} f_{s,i} =& (1+z_s)10^{-0.4\mu_s} \times 10^{0.4Z_{s,i}} \times \int_{\lambda_r^{\min}}^{\lambda_r^{\max}} S_s(t_s^i,\lambda_r) \\ & \times 10^{-0.4A^s_{\mathrm{MW}}\xi(\lambda_R[1+z_s];R_{\mathrm{MW}})} \times \mathbb{B}_{s,i}(\lambda_r[1+z_s])\lambda_r d\lambda_r \end{aligned} \tag{3.2}$$

where $\mathbb{B}$ denotes the bandpass transmission functions, μ_s is the distance to the supernova, $Z = 27.5$, and z_s denotes the supernova redshift.

Using Equations 3.1 and 3.2 and the procedure outlined in Figure 3.2, a light curve is generated based on the input parameters that can then be optimized.

3.3 Generative model implementation details

Our implementation of the BayeSN model includes four parameters to be fit for each supernova s: the distance parameter μ_s, the shape parameter θ_1^s, the dust parameter A_V^s and the residual parameter ν_s, which goes into the calculation of the $\mathbf{E_s}$ matrix.

We initially model the SED as a cubic spline. The integration of the SED is performed by evaluating the surface at a given set of linearly spaced time and wavelength points, or “knots". We implement Equations 3.1 and 3.2 using this cubic spline representation of the

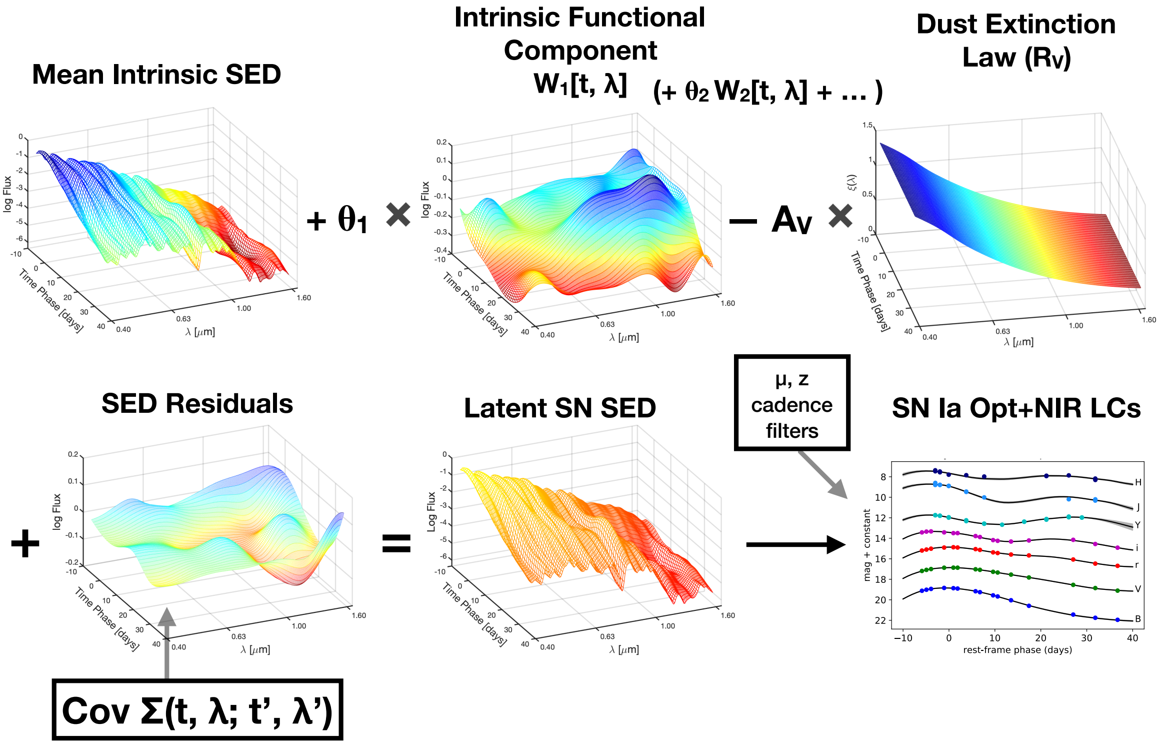

Fig. 3.2 A schematic of the BayeSN forward generative model for the light curve of a single supernova (Mandel et al., 2022).

functional parameters, and approximate the passband integral (seen in Equation 3.2) using a Riemann sum.

In this implementation, only the first principal component is used, and the intrinsic SED from Equation 3.1 is represented as

$$\mathbf{W} = \mathbf{W_0} + \theta_1^s \mathbf{W_1} + \mathbf{E_s} \tag{3.3}$$

where each term is a spline representation of a term in Equation 3.1. The matrix $\mathbf{W}$ describes the log of the intrinsic SED $S_s(t, \lambda_r)$, and is a grid of spline knots (of shape $N_\lambda \times N_t$) in time-wavelength space. The $\mathbf{W_0}$ and $\mathbf{W_1}$ matrices used in this implementation represent the functional principal components, and the $\mathbf{E_s}$ matrix is a representation of the residual surface $\varepsilon(t, \lambda)$. This aims to capture any residual variance that is not represented by the functional principal components.

The precomputed, parameter independent matrix $\mathbf{J_t}$ performs a transformation from the value of the SED at the time knot locations t to any arbitrary set of time points $\mathbf{t}^*$ (Press et al., 2007). $\mathbf{J_t}$ is of dimension $N_t \times N_{t^*}$ and is solely a function of the the existing time knots t

and desired time points t^*:

$$\mathrm{SED}|_{\mathbf{t}^*} = \mathbf{J}_\mathbf{t}^\mathbf{T}\mathrm{SED}|_\mathbf{t} \tag{3.4}$$

The corresponding matrix $\mathbf{J}_\lambda$ is of shape $N_{\lambda^*} \times N_\lambda$ performs a similar transformation in wavelength:

$$\mathrm{SED}|_{\lambda^*} = \mathbf{J}_\lambda\mathrm{SED}|_\lambda \tag{3.5}$$

Once the matrix of spline knots defining the intrinsic SED $\mathbf{W}$ is created, the $\mathbf{J}_\mathbf{t}$ and $\mathbf{J}_\lambda$ evaluate it over a densely sampled grid of times and wavelengths:

$$\mathbf{W}^* = \mathbf{J}_\lambda\mathbf{W}\mathbf{J}_\mathbf{t}^\mathbf{T} \tag{3.6}$$

This dense grid is of shape $N_{\lambda^*} \times N_{t^*}$ and will be used to evaluate the passband integrals over time .

To create the apparent observed SED, we must include the effects of dust. We use the dust extinction law from Fitzpatrick (1999), implemented in the `extinction` Python package, to describe the modification to the intrinsic SED. The observed SED can be calculated as:

$$\tilde{\mathbf{S}} = \mathbf{S_0}\exp(-\gamma(\mathbf{W}^* + A_v\Xi)) \tag{3.7}$$

where $\mathbf{S_0}$ is the Hsiao template evaluated at $\mathbf{t}^*$ and λ^*, $\gamma = \frac{\log(10)}{2.5}$, A_v is the dust extinction parameter, and Ξ is a matrix containing the Fitzpatrick (1999) extinction law evaluated at each wavelength point, tiled to form the shape $N_\lambda \times N_t$.

To generate a light curve, we use Equation 3.2 to calculate the flux profile $\mathbf{f}$ (evaluated at times $\mathbf{t}^*$) in each wavelength filter i:

$$\log_{10}(\mathbf{f}_i) = 0.4(\mathrm{ZPT} - \mu_s - M_0 - \delta M)\mathbf{h}\tilde{\mathbf{S}} \tag{3.8}$$

where ZPT and M_0 are constants, δM is set to zero, $\mathbf{h}$ applies the effects of observing through our Milky Way galaxy and also performs the integration under the transmission function as described in Equations 2.4 and 3.2, and $\tilde{\mathbf{S}}$ is the observed SED calculated in Equation 3.7.

3.4 Description of single-supernova Pyro model

For a single supernova, there are 27 trainable parameters: μ, A_v, θ, and 24 elements of the $\mathbf{E_s}$ matrix. $\mathbf{E_s}$ aims to capture any residual variance in the SED and is drawn from a normal prior $\mathrm{N}(0,\Sigma)$, where Σ is a predetermined residual covariance matrix estimated by Thorp et al. (2021). This can be reparameterized with a vector ν drawn from a uniform Multivariate

Normal distribution, which is then multiplied by the Cholesky decomposition of Σ. Each parameter has a corresponding prior distribution[1] :

$$\mu \sim \mathrm{N}(\hat{\mu}, 100) \tag{3.9}$$

$$\theta \sim \mathrm{N}(0,1) \tag{3.10}$$

$$A_V \sim \mathrm{Exp}(1/\tau) \tag{3.11}$$

$$v_i \sim \mathrm{N}(0,1),\ i = 1, ..., 24 \tag{3.12}$$

where $\tau = 0.194$, a population mean A_V that is determined from inference of the population-level parameters (Mandel et al., 2022; Thorp et al., 2021). $\hat{\mu}$ is a distance estimate obtained independently of the light curve data, calculated from the observed redshift using the `cosmo.distmod()` function from the `astropy` library to evaluate Equation 2.6, using cosmological parameters from Riess et al. (2016) (Astropy Collaboration et al., 2013, 2018). The standard deviation of 10 on the distribution of μ (where $\hat{\mu}$ is usually between 34 and 38) approximates an essentially uninformative prior. The individual values of v_i are reshaped into a matrix of shape (6,4), which is then padded with zeros on the top and bottom to create the $\mathbf{E}$ matrix. We also include a measurement error σ_f that relates observed flux to model flux (calculated using Equation 3.8) via:

$$\mathbf{f}_{\mathrm{obs}} \sim \mathrm{N}(\mathbf{f}, \sigma_f) \tag{3.13}$$

The overall posterior for a single supernova is as follows:

$$p(\mu, \theta, A_V, v | \mathbf{f}_{\mathrm{obs}}, \hat{\mathbf{H}}) = p(\mathbf{f}_{\mathrm{obs}} | \mu, \theta, A_V, v, \sigma_f, \hat{\mathbf{H}})\, p(\theta)\, p(A_V|\tau)\, p(\mu)\, p(v|\mathbf{L}_\Sigma) \tag{3.14}$$

where $\hat{\mathbf{H}}$ represents the model hyperparameters $\{R_V, \tau, \mathbf{L}_\Sigma, \mathbf{W_0}$, and $\mathbf{W_1}\}$

In our VI implementation, we use a Multivariate Normal surrogate posterior and fit values of the parameters μ, θ, v and A_V such that the KL Divergence between the surrogate posterior and the true posterior in Equation 3.14 is minimized.

Figure 3.3 shows a probabilistic graphical model of the BayeSN model for a single supernova. The generated flux is conditional on μ_s, A_v, θ, and v, and is measured in each wavelength band (g, r, i, and z) and at each observed time point, denoted by the plates labeled “bands" and “observations".

The Pyro model initially samples values from the prior distribution and generates light curves with those values, using the process described in Section 3.3. Pyro’s `SVI` infrastructure

[1] Here we use the notation $\mathrm{N}(\mu, \sigma^2)$ for normal distributions, where μ denotes mean and σ^2 denotes variance, and $\mathrm{Exp}(1/\tau)$ for an exponential distribution with rate parameter τ

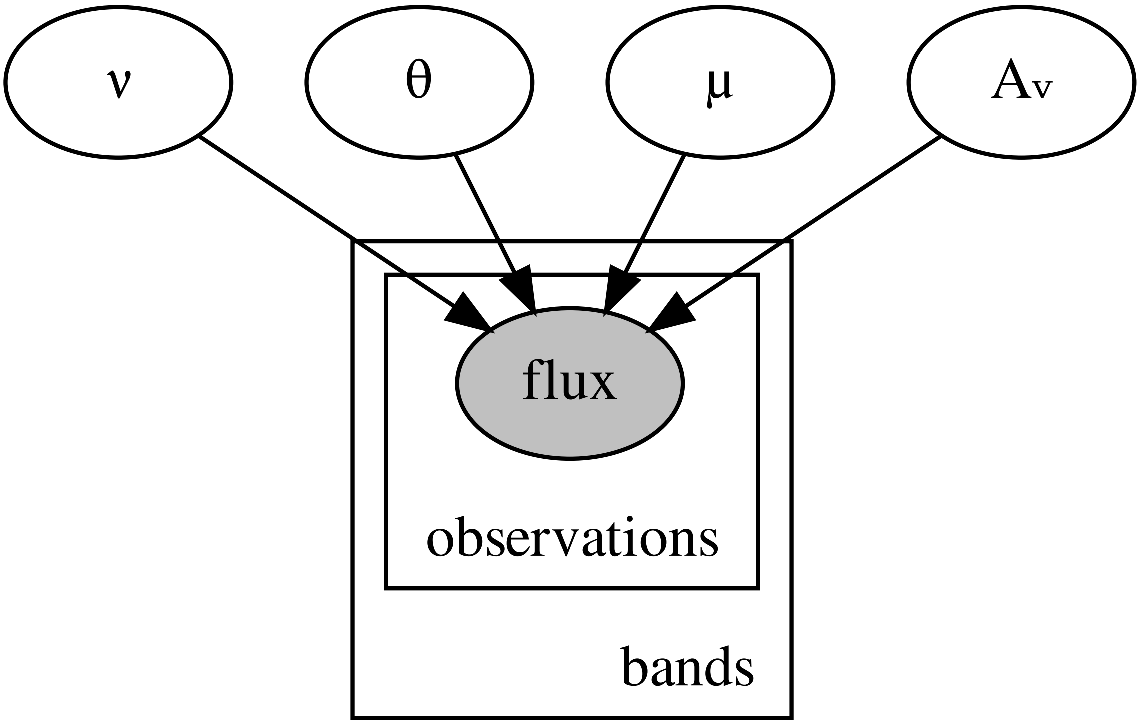

Fig. 3.3 A probabilistic graphical model of the BayeSN-VI model, implemented in Pyro.

allows us to calculate the ELBO conditional on the observed data, and the posterior estimates of the parameters are updated using a specified optimizer. The model continues updating the parameter values to minimize the negative ELBO for a given number of iterations.

The Adam optimizer, which exponentially decreases the moments of gradients to adaptively fit different parameters, was used for this and all other Pyro models in this work (Kingma & Ba, 2014). Additionally, gradient clipping provided additional smoothing in this stochastic optimization.

We found that using a Multivariate Gaussian (with the `AutoMultivariateNormal` guide object in Pyro) to approximate the posterior led to the best results because its full-rank covariance matrix allowed it to model even weak correlations between parameters, which is key to ensure the that uncertainties in the light curve shape and levels of dust extinction are correctly propagated to the SED and distance.

The default initialization of parameter values is to the median of the prior distributions. We initialized using the Laplace approximation of the posterior (see Appendix 1 for a derivation of the Laplace Approximation) as this produced more reliable results than Pyro's default initialization (using the prior median with a standard deviation of 0.1 for all parameters). In this case, we used 3,000 iterations to fit the Laplace Approximation.

The `AutoLaplaceApproximation` guide in Pyro calculates the MAP estimate, and uses the Hessian to calculate a covariance matrix for a Multivariate Gaussian distribution. This distribution is then fine-tuned using 6,000-10,000 steps with the `AutoMultivariateNormal` guide.

Chapter 4

VI model results on simulated type Ia supernova light curves

4.1 Method to simulate light curves

Once the generative BayeSN model for a single supernova has been implemented in Python, it can be used to generate any number of simulated supernova light curves with predetermined input parameter values. Once this simulated population of supernovae is created, we can examine the accuracy of the VI light curve fitting model by comparing the parameter values determine by VI to the original values used to as input for the simulated light curve. This method allows us to assess systematic uncertainties and biases of the VI model as well as compare the performance of other methods, such as MCMC, against VI and the ground truth values.

To create a simulated light curve, a synthetic redshift z was generated from $\mathrm{U}(0.015, 0.08)$, the range of redshifts targeted by the Foundation Supernova Survey, whose data we will be analyzing later in this work (Foley et al., 2018; Jones et al., 2019). The true distance μ_{mean} was calculated using the realtion in Equation 2.6 as described above. Synthetic observed time points were generated using a maximum time of B-band brightness t_{max} randomly sampled from a uniform distribution between 57100 and 57800 MJD, and the earliest observation time sampled from between 5 and 10 days before the maximum time. Time points were generated every six days after this randomly chosen first observation, and scaled to the rest frame of the supernova as follows:

$$t = \frac{T - t_{max}}{1 + z} \tag{4.1}$$

where z is the redshift. The “true" value of μ_s was calculated from this z value, and “true" values of θ,A_v and ν were randomly sampled from the distributions listed above, with the exception of θ, which was sampled from U(-1.33, 2.78). Using these time points and parameter values, synthetic light curves were generated using the process outlined in Section 3.3.

These generated light curves were used as the “observations" passed into the Pyro model. None of the “truth" values were used in the Pyro model, which was initialized as usual from random draws from the prior distributions. Simulated observational errors of 2% were used for the generated light curves, in line with typical photometric uncertainties in the Foundation dataset. The model was then trained as usual, calculating the log probability of the parameters conditioned on the generated “observed" light curves, and optimizing the model parameters of the approximate model of the posterior accordingly.

Figure 4.1 shows a simulated supernova light curve and its fit using the VI model. The points in each wavelength band represent the “observed" values, while the lines represent the fit from the model. To create and visualize a credible interval on the surface, the predictive posterior was used to generate 100 potential curves, the standard deviation of which is seen in the shaded areas in Figure 4.1.

Using this scheme, a simulated population of 150 supernovae were generated to analyze the performance of the VI model.

4.2 Performance on simulated data

4.2.1 Single simulated supernova

Figure 4.2 shows the estimated posterior distributions of the three single-value parameters (μ, θ, and A_v) for a simulated light curve. The simulated supernovae were fit using using VI initialized from the prior median, VI initialized from the Laplace approximation, and MCMC. Generally, MCMC results are seen as the gold standard estimate of the posterior distribution to compare against, but in this simulated case, the real “true" values of the parameters can be compared with the output distributions of approximate methods.

Overall, the VI, VI + Laplace, and MCMC parameter distributions agree quite well with the true values, with the mode of the posterior, as estimated by MCMC, agreeing closely with the true values. Note that the 1σ contours for the VI and VI + Laplace distributions are much smoother than MCMC because they are modeled as Multivariate Gaussians, while no such constraint is placed on the MCMC results. This also gives the added benefit that VI returns an approximate posterior distribution with a tractable form that can be arbitrarily sampled,

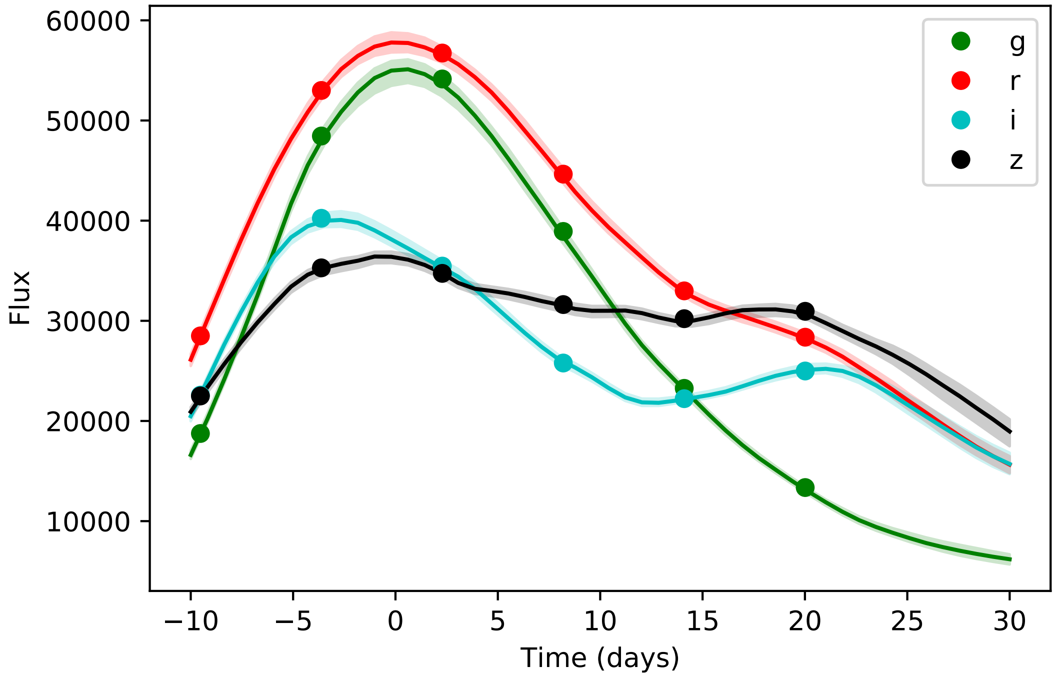

Fig. 4.1 A simulated supernova light curve. Points denote the synthetic “observed" flux values, while the lines denote the fit generated using VI. Shaded regions denote 1σ uncertainties.

while MCMC returns only a certain specified number of samples. Both VI distributions also approximate the spread of the MCMC distribution reasonably well, indicating that the Pyro model is not over- or under-estimating uncertainties on parameters relative to the MCMC distribution.

4.2.2 Simulated supernova population

To further assess the VI model implemented in Pyro, we simulate a population of 150 supernovae with random “true" parameter values drawn as described previously. Each simulated supernova was then fit with the Pyro model using VI initialized with the Laplace Approximation with a Multivariate Gaussian surrogate posterior, which returned a mean and variance for each parameter. This analysis allowed us to explore systematic biases and the accuracy of the estimation of the parameter values and their uncertainties.

Figure 4.3 shows a comparison between the “true" values for μ, θ, and A_V, and the values determined from the VI fit, as well as the residuals (fit value - true value) for each parameter. Generally, good agreement is seen between the true values and those from the

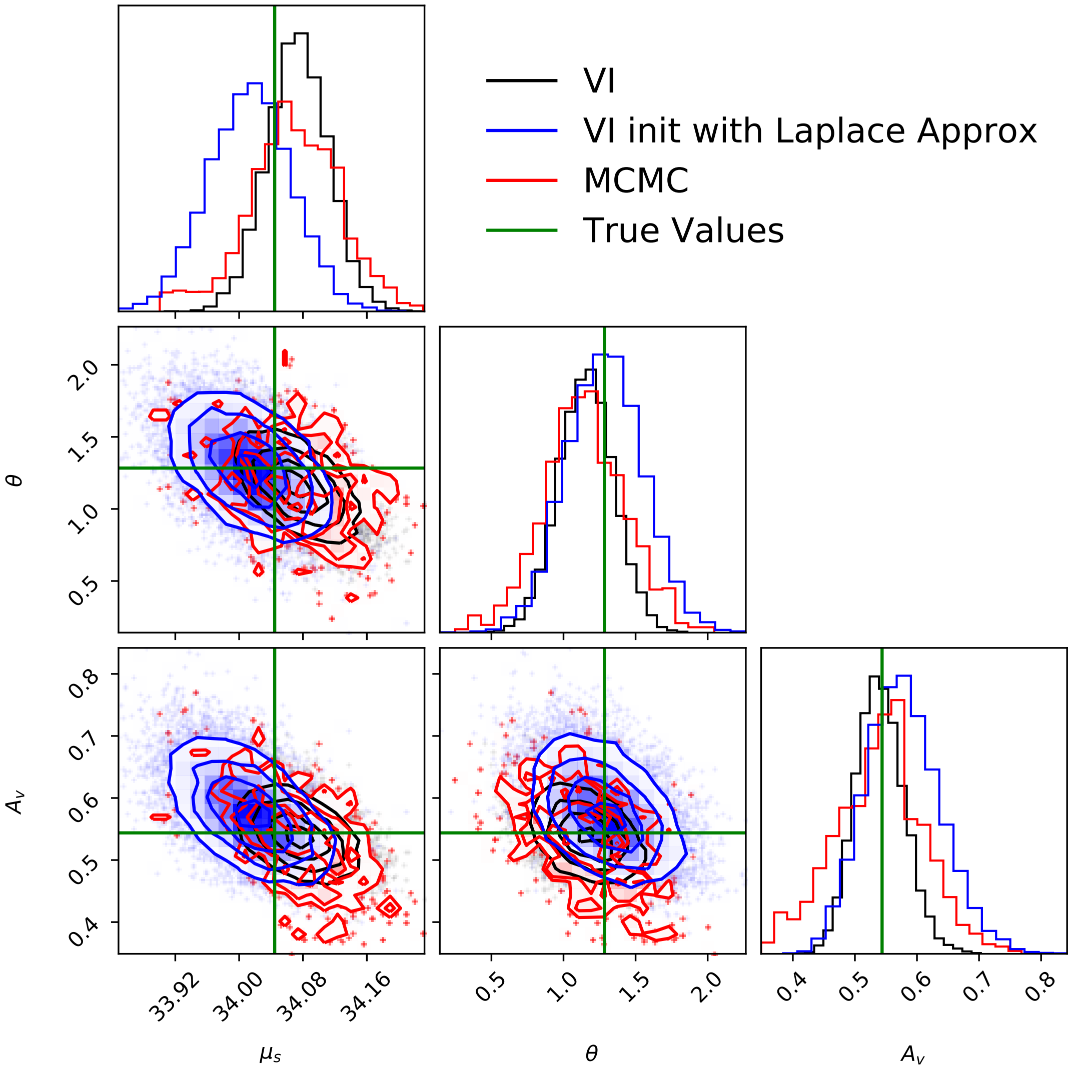

Fig. 4.2 Corner plot comparing the distributions of μ, θ, and A_V fit with VI, VI initialized on the Laplace Approximation, and MCMC, with the true parameter values used to initialize the simulation in green. Distributions of each combination of two parameters are plotted, and 1D histograms of each parameter are shown along the diagonal.

VI across all parameters. There is no obvious bias in μ, and the VI model appears to fit the distance extremely well, with a maximum residual value of 0.175 and a standard deviation of 0.05. In θ and A_V there appears to be a slight bias as a function of the parameter value, with the residual decreasing as the value of the parameter increases, which could lead to subtle biases in estimation of cosmological parameters based on these values. The dust extinction parameter A_V is based on an exponential prior, so Pyro fits the multivariate

Gaussian surrogate posterior to $\log A_V$, since it is constrained to be positive. This distribution cannot approximate a distribution that peaks at zero, thus systematically overestimating the value of A_V for supernovae with intrinsically low A_V. Different true values of A_V can require differently shaped approximate distributions. For A_V close to zero, a good posterior approximation would be asymmetric and peak close to zero while staying positive, while for a larger value of A_V, the posterior approximation would more closely resemble a Gaussian. In general, approximating A_V well is a difficult task, as is discussed in further detail in Chapter 6.

In addition to comparing the true and mean fit values of the distance parameter μ, we also assessed the variances returned from the VI fit. 82% of the simulated population had the true value within 1σ of the fit value, while 97.5% had the true value within 2σ. Since these values would be expected to be closer to 68% and 95%, this suggests that the VI model may be slightly underestimating the uncertainties of the fit parameters. That being said, MCMC is also not guaranteed to accurately estimate parameter uncertainties, and in Figure 4.2 the width of the MCMC and VI distributions appear to be similar. Overall, these biases could merit additional analysis and further consideration when using the VI model for precision cosmology, but the current performance is overall quite promising.

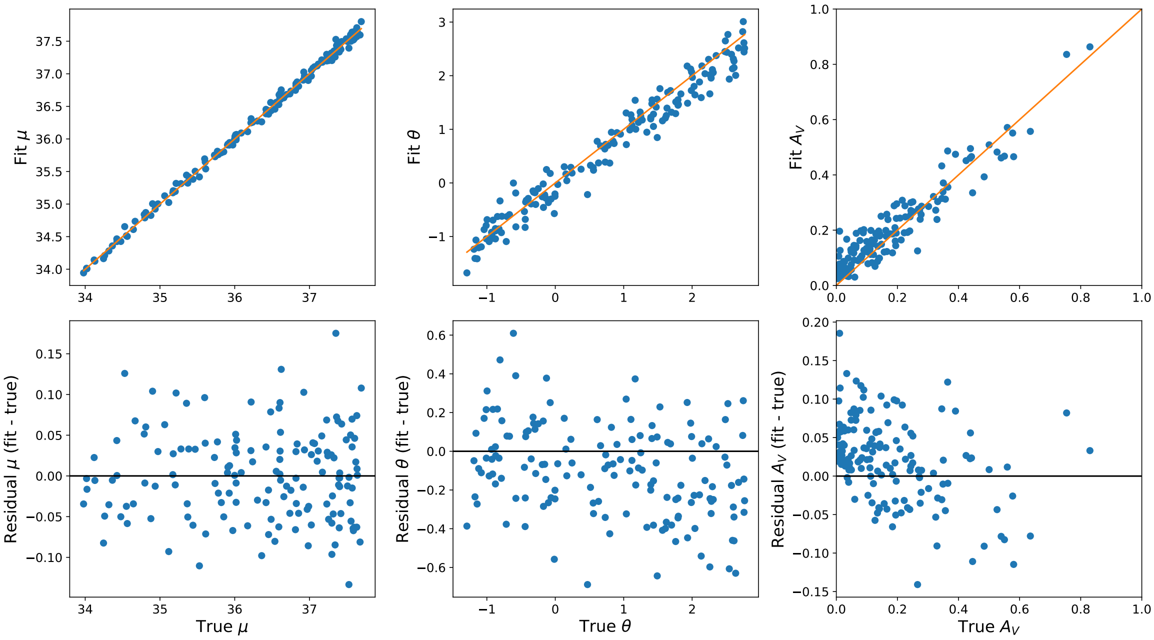

Fig. 4.3 Comparison of true vs. fit values of μ, θ, and A_V for 150 simulated supernovae. Top row plots true vs. fit values with a line at $y = x$ for comparison, bottom row plots residuals (fit - true) as a function of the true parameter values.

Figure 4.4 shows the residuals (fit - true) for μ, θ, and A_V for this simulated population plotted against each other. A negative correlation can be seen between the residuals in μ vs. A_V and μ vs. θ, which suggests a tradeoff between these values in estimating the overall SED. There is no significant correlation between the residuals in θ vs. A_V, which makes sense because the shape of intrinsic SED and the effects of dust are not directly related, but both impact the observed flux.

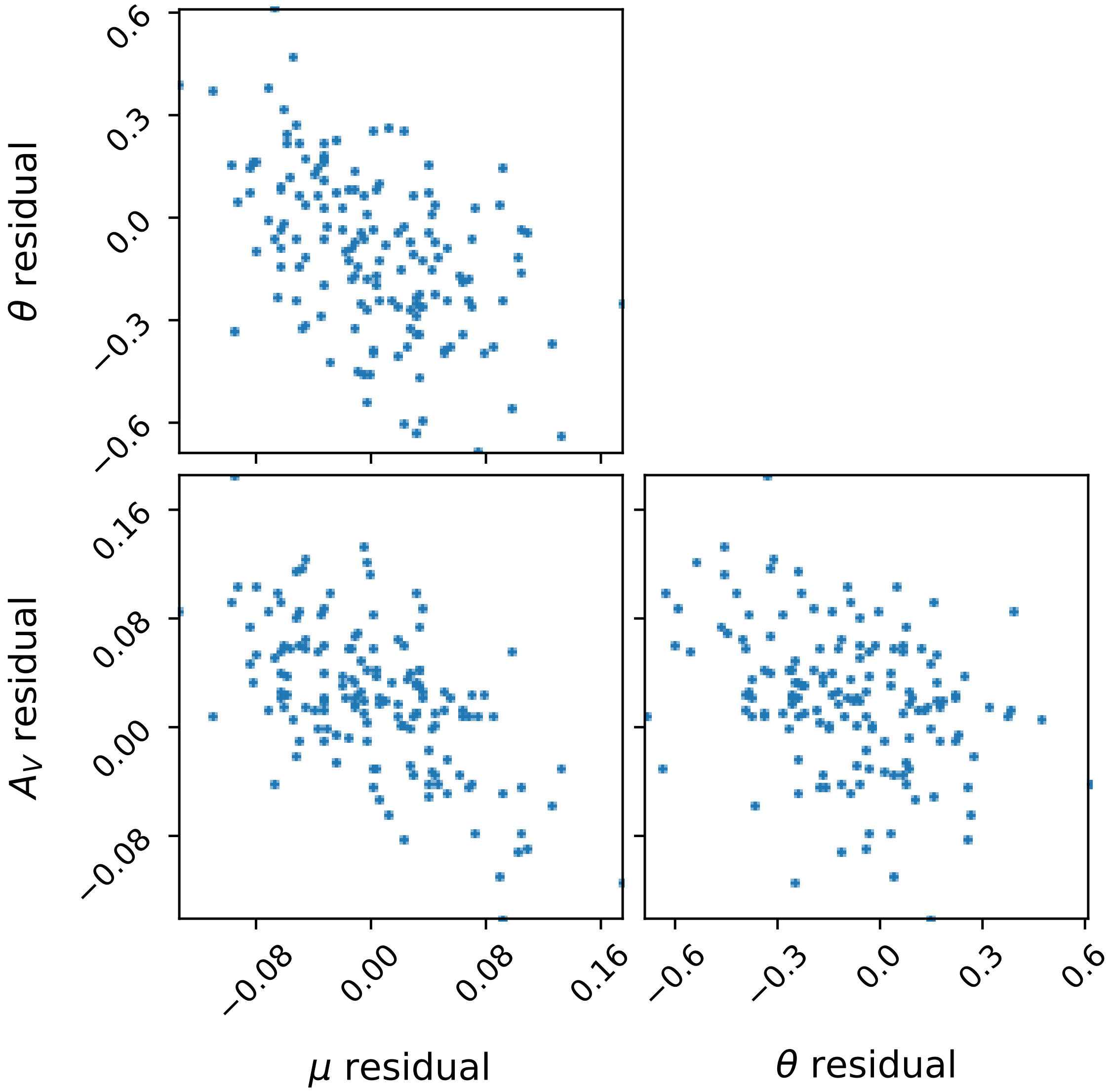

Fig. 4.4 “Triangle" plot of each combination of μ, θ, and A_V residuals (fit - true) for 150 simulated supernovae. These are essentially the lower row of Figure 4.3 plotted against each other.

Chapter 5

Results on Foundation Dataset & Comparison with MCMC

5.1 Overview of dataset

The Foundation dataset is a set of 157 Type Ia supernovae with well-sampled light curves observed with the Pan-STARRS telescope located in Maui, Hawaii (Foley et al., 2018; Jones et al., 2019). Observations were made in the g, r, i, and z bands, and specifically targeted supernovae with low redshifts (Foley et al., 2018).

5.2 Methods

Real observed supernova data created two additional complications that were not present in previously described simulations: irregular sampling and Milky Way dust extinction. The dataset contained sampling times and sampled fluxes at each time for each wavelength band, but a supernova could be sampled at different times for each wavelength band, and some bands could have more observations than others for the same supernova. For each supernova, we were also given the observed redshift z as well as the CMB redshift z_{CMB} (which are corrected for peculiar velocity) and Milky Way reddening $E(B-V)_{\mathrm{MW}}$.

The same single-supernova Pyro model described in Section 3.4 was used to fit each supernova's light curve individually, with ν, θ, μ, and A_V being fit parameters. $\hat{\mu}$ was calculated using the CMB redshift function as previously described. Each supernova was fit for 3000 steps to fit the Laplace Approximation and, was fit for a further 6000 steps using a Multivariate Normal guide. As before, a Multivariate Normal surrogate posterior yields a mean and variance for each fitted parameter.

Each supernova was fitted in less than a minute, approximately one tenth of the time used to fit the same supernova using MCMC. However, runtimes using MCMC vary widely depending on the number of chains used and the desired number of samples, and could be significantly longer.

5.3 Comparison of results to expected Hubble relation

Figure 5.1 shows a “Hubble diagram" of the fit distance modulus μ vs. observed redshift z for the Foundation dataset. The black line denotes the expected Hubble relation using an assumed $H_0 = 73.24$ and $\Omega_0 = 0.28$, with these values derived from Riess et al. (2016). The fit distance values agree well with the black Hubble relation line, indicating a good fit from the VI model. The root mean squared error (RMSE) for these fit distance values was calculated to be 0.123, consistent with the value from Thorp et al. (2021) for the same dataset using HMC in Stan.

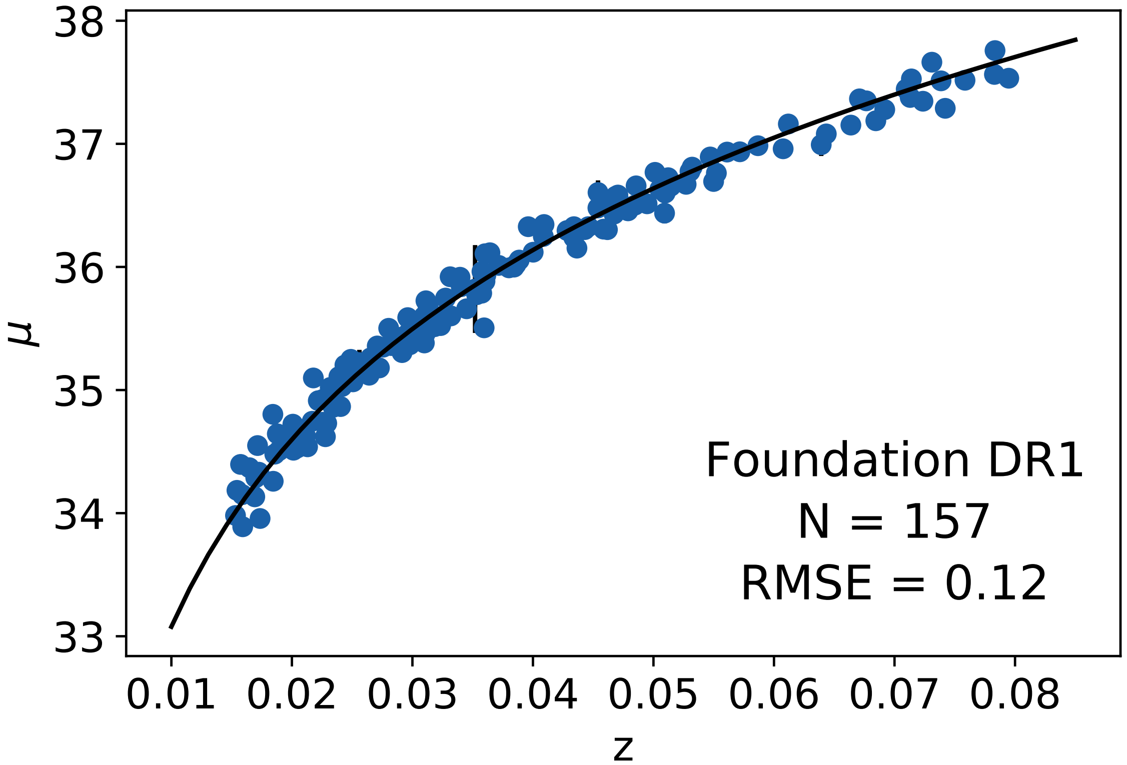

Fig. 5.1 “Hubble diagram" of distance modulus μ vs. redshift z for 157 supernovae the Foundation dataset. Black line in indicates expected Hubble relation with $H_0 = 73.24$ and $\Omega_0 = 0.28$ (Riess et al., 2016). Error bars indicate 1σ for each supernova. The RMSE for this dataset is 0.12.

Figure 5.2 shows the “Hubble residuals" for the Foundation dataset, essentially the difference between the points and the black Hubble relation line in Figure 5.1. The dotted lines represent the peculiar velocity error “envelope", which comes from an assumed uncertainty on the peculiar velocity correction of 150 km/s. At lower redshifts, this is a larger fractional uncertainty, leading to the shrinking effect over redshift. The Hubble residuals show a slight negative bias, likely due to bias in A_V that is explored further in the residual plot for the μ and A_V parameter (Figures 5.4 and 5.3).

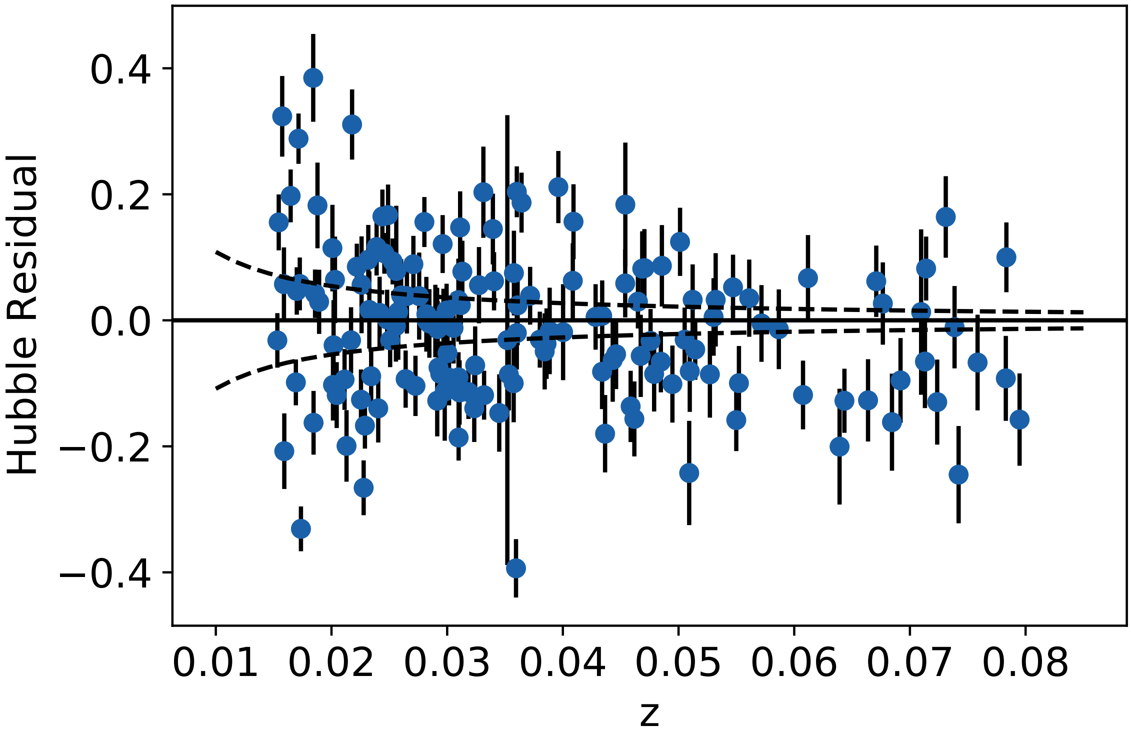

Fig. 5.2 “Hubble residuals" (difference between fitted μ values and expected Hubble relation shown in Figure 5.1) for 157 supernovae in Foundation dataset. Black horizontal line at zero is for reference, while dotted black curves represent the peculiar velocity uncertainty envelope. Error bars represent 1σ.

5.4 Residuals analysis

In Section 4, we evaluated the performance of the VI model on simulated supernova light curves by comparing the posterior means for each fit parameter to the “true" values used in the simulations. Here, we used VI to fit the light curves of real supernovae from the Foundation dataset, and thus do not have ground truth values to compare our variational posterior to.

However, Thorp et al. (2021) fit the Foundation dataset using MCMC in Stan, and we can use these results (which are the current “gold standard") as a basis for comparison.

Figures 5.3, 5.4, and 5.5 show the residuals for A_V,μ, and θ, respectively, for the Foundation dataset. Residuals were calculated as VI - MCMC and are shown as a function of the MCMC parameter value.

Figure 5.3 shows a tendency to overestimate A_V for lower values, similar to what was seen in Figure 4.3 for the simulated supernova population. This is likely for the same reason previously described, that approximating an exponential distribution with a log-Normal overestimates low values of A_V. But, unlike in the simulation-based results, there is a more significant underestimation of A_V relative to MCMC for higher A_V values. This could be due to biases in the VI model, or indication that the MCMC fit may be overestimating certain value of A_V.

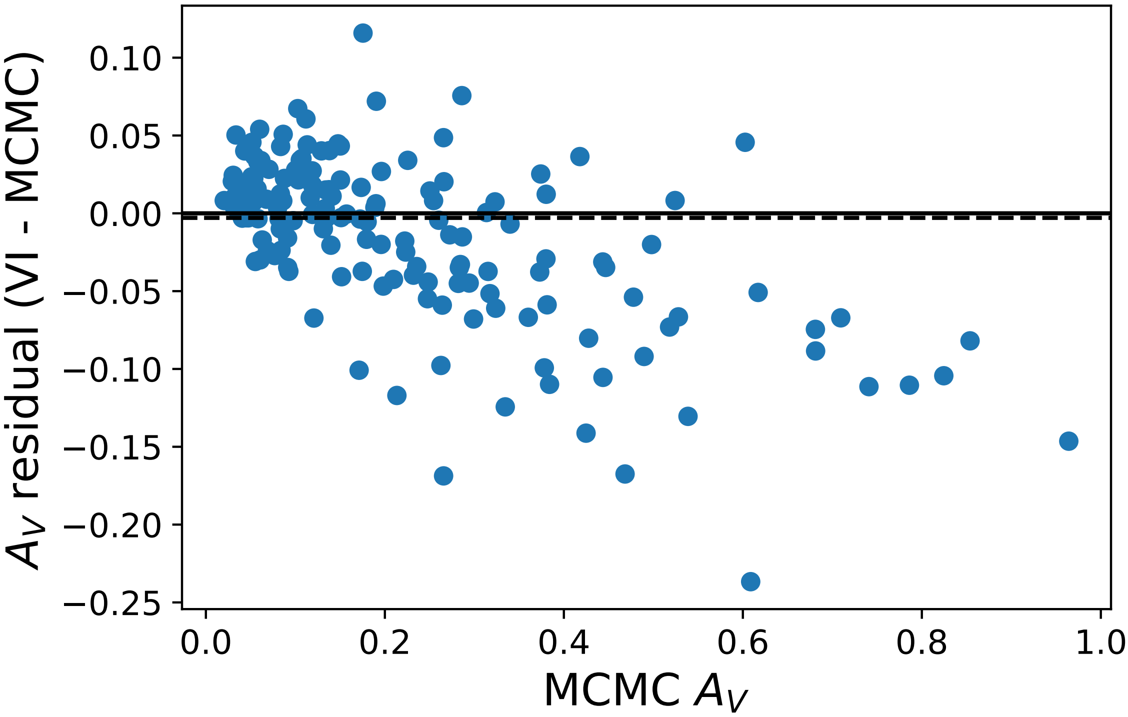

Fig. 5.3 Residual plot of MCMC A_V value vs. A_V residual (VI - MCMC) for 157 supernovae in the Foundation dataset. A solid horizontal line at zero is provided for reference, and the dashed horizontal line denotes the median residual value for the dataset.

There is also a slight negative bias in μ. Figure 5.4 shows the residuals in μ as a function of the MCMC values of μ, but colored by the values of A_V for each supernova. A clear trend is visible; higher values of A_V are overestimated by the VI model while lower values of A_V are underestimated.

This trend could be caused by certain selection effects. Supernovae with higher values of μ are less likely to be observed with high values of A_V, a trend which is reflected in the Foundation dataset. This is expected, because higher μ means that an object is farther away, which means that it is less likely to be observed if it has high levels of dust extinction (A_V) dimming its brightness. The resulting effect on the data is that the VI model will underestimate μ for a given value of A_V, leading to the trend seen in Figure 5.4.

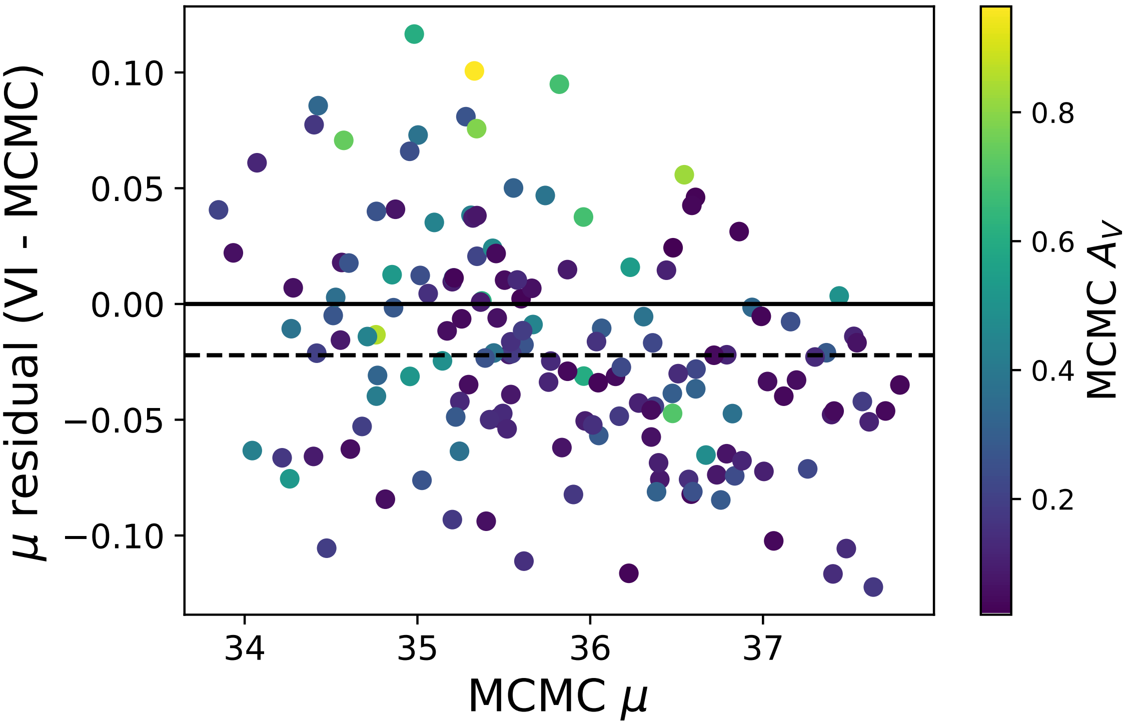

Fig. 5.4 Residual plot of MCMC μ value vs. μ residual (VI - MCMC) for 157 supernovae in the Foundation dataset. Points are colored by the corresponding value of A_V. A solid horizontal line at zero is provided for reference, and the dashed horizontal line denotes the median residual value for the dataset.

Figure 5.6 explores the correlations between the residuals in μ, θ, and A_V. Many of the same relationships seen in the simulated dataset (Figure 4.4) are reflected in the real data, such as negative correlations between the residuals in θ vs. μ and A_V vs. μ. The correlation between A_V vs. μ is the strongest, which can be explained by the trends in the color in Figure 5.4 and the previously discussed potential selection effects. Supernovae with higher μ residuals tend to have higher values of A_V , and A_V is shown to be negatively biased (see Figure 5.3). Thus, there would be a negative correlation between the μ and A_V residuals, as seen in Figure 4.4.

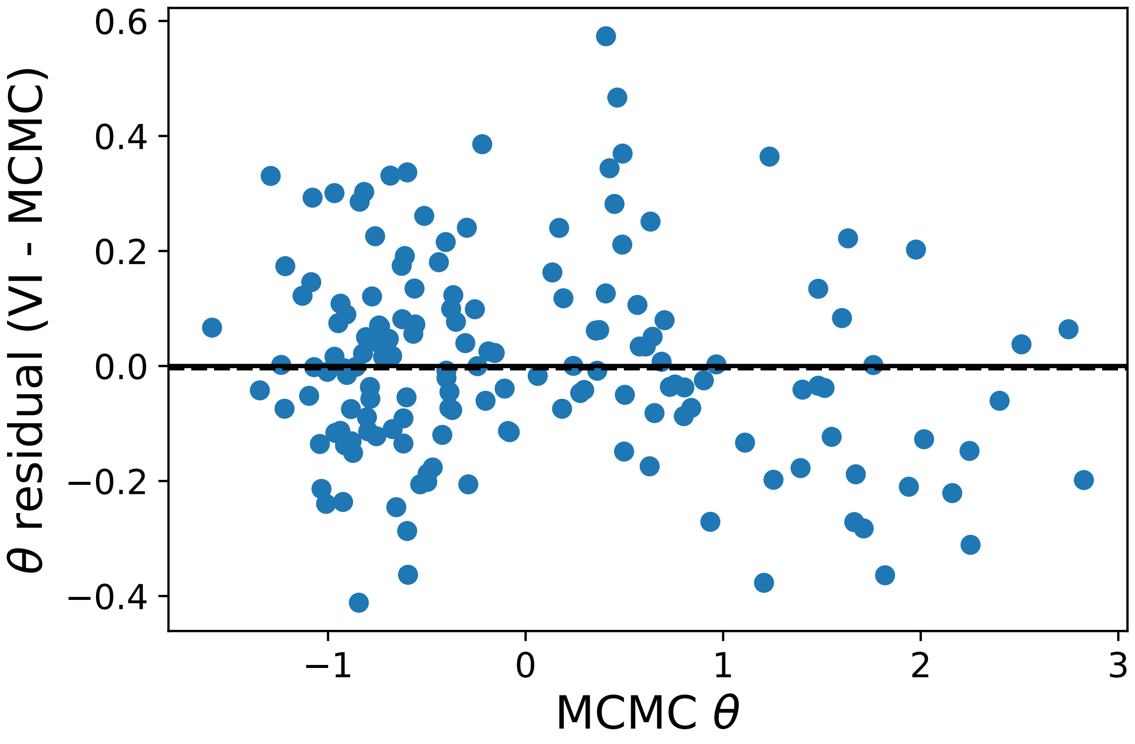

Fig. 5.5 Residual plot of MCMC θ value vs. θ residual (VI - MCMC) for 157 supernovae in the Foundation dataset. A solid horizontal line at zero is provided for reference, and the dashed horizontal line denotes the median residual value for the dataset.

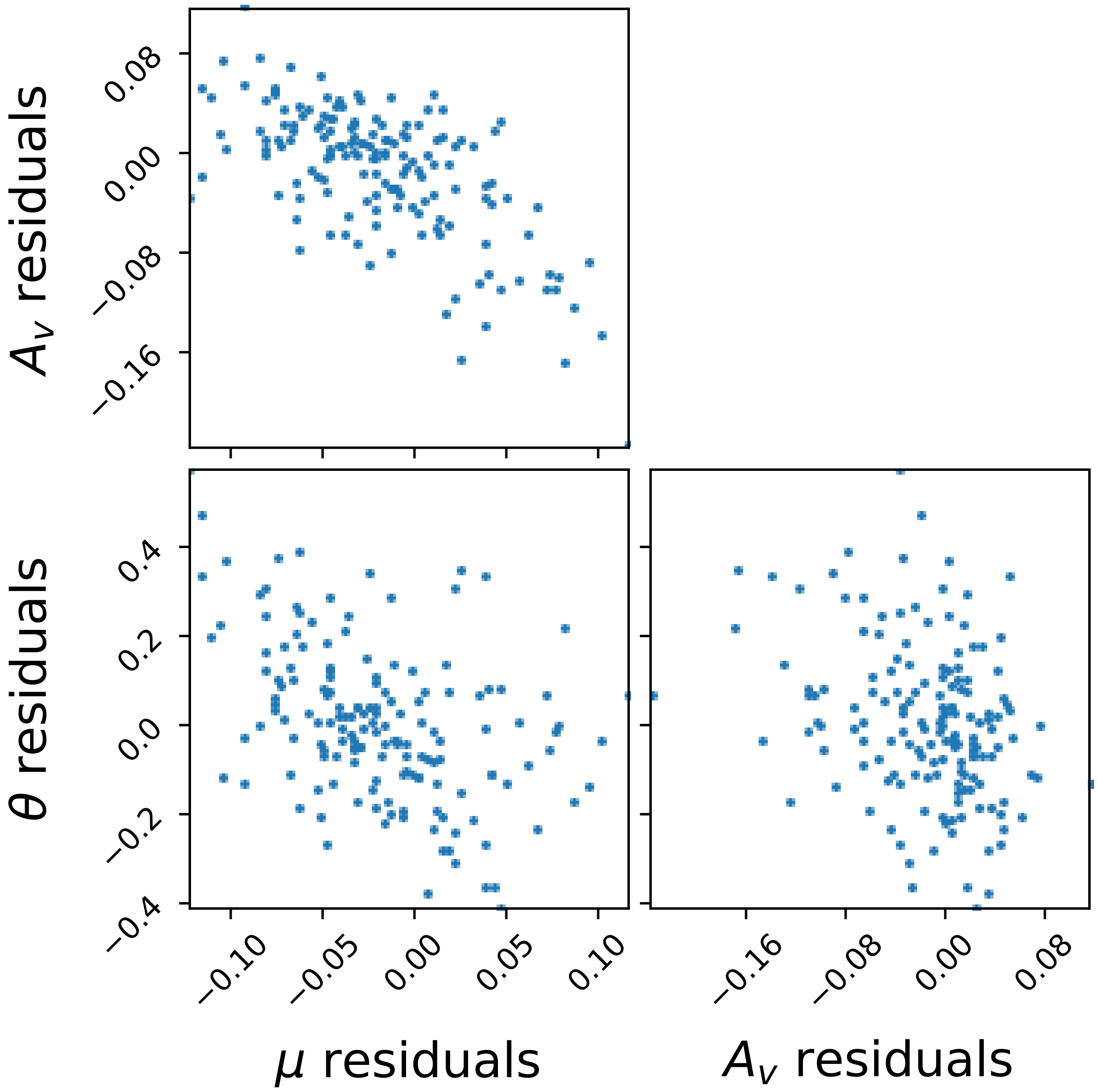

Fig. 5.6 “Triangle" plot of each combination of μ, θ, and A_V residuals (VI - MCMC) for Type Ia SNe in the Foundation dataset. This is essentially Figures 5.3, 5.4, and 5.5 plotted against each other.

Chapter 6

Investigating & mitigating bias in dust extinction

6.1 Motivation

In Chapters 4 and 5, we fitted the light curves of simulated and real Type Ia supernova datasets using a VI model implemented in Pyro. We then compared the results from the VI fit to the input parameters for the simulated population, and to the MCMC fit for the Foundation dataset. In both cases, we calculated and analyzed the residuals (fit − true) of μ, θ, and A_V. Here, we found a trend of A_V residual being systematically overestimated for values of A_V close to zero (see Figures 4.3 and 5.3). While it is true that the posterior mean will overestimate the true value of A_V at these values (since A_V is constrained to be positive), it seemed that our current method was not modeling the posterior distribution of A_V as well as it could be. As such, we investigate alternate ways to model A_V that may address this bias. We specifically aim to improve performance as $A_V \to 0$, while maintaining the physically informed exponential prior for the rest of the dataset with larger values of A_V.

6.2 Softplus transformation

6.2.1 Theoretical overview

To model a constrained variable, Pyro models an unconstrained variable and applies a constraining transformation to the unconstrained variable's posterior distribution. Previously, we have drawn A_V from an exponential prior distribution:

$$A_V \sim \text{Exp}(1/\tau) \tag{6.1}$$

where τ is a population level parameter which Thorp et al. (2021) determined to be 0.194.

Pyro models the posterior as a log-normal distribution, where

$$\log(A_V)|\mathbf{f}_{\text{obs}} \sim N(\mu, \sigma^2) \tag{6.2}$$

for some mean μ and variance σ^2. As discussed in previous chapters, a log-normal variational distribution cannot model a distribution that peaks at or near zero, and thus leads to the VI model overestimating A_V for supernovae with intrinsic A_V close to zero. For these low A_V values, a variational distribution more like a truncated Gaussian would provide a better posterior approximation.

There are various transformations and parameterizations that could be applied to circumvent this log-normal modeling. A common solution in machine learning is using the Recified Linear Unit (ReLU) function to transform an unbounded input into a positive output (Agarap, 2018). The softplus function is a differentiable version of ReLU that performs the following transformation:

$$\text{Softplus}(x) = \log(1 + \exp(kx))/k \tag{6.3}$$

where k is a sharpness parameter. By defining a new variable x such that $A_V = \text{Softplus}(x)$, we can enforce a positive constraint on A_V while also having a distribution that peaks at zero. This more sophisticated transformation maps a Gaussian-distributed unconstrained variable to a distribution closely resembling a truncated Gaussian, in theory allowing for more accurate fitting of low A_V values.

Using this transformation, the prior distribution on x is:

$$\begin{aligned} P(x) &= \frac{1}{\tau} \exp\left(\frac{-A_V(x)}{\tau}\right) \times \frac{1}{1 + \exp(-kx)} \qquad (6.4) \\ &= \frac{1}{\tau(1 + \exp(-kx))} \exp\left(-\frac{\log(\exp(kx) + 1)}{k\tau}\right) \qquad (6.5) \end{aligned}$$

which can be seen in Figure 6.1. When transformed via a softplus function, this leads to an exponential prior on A_V.

6.2.2 Implementation

The single-supernova Pyro model discussed in Chapters 4 and 5 was modified to sample a new parameter x from the prior distribution outlined in Equation 6.5 with a value of $k = 25$. To sample from this prior, a custom distribution object was created in Pyro by extending the PyTorch `TorchDistribution` base class and overriding the sampling and log probability

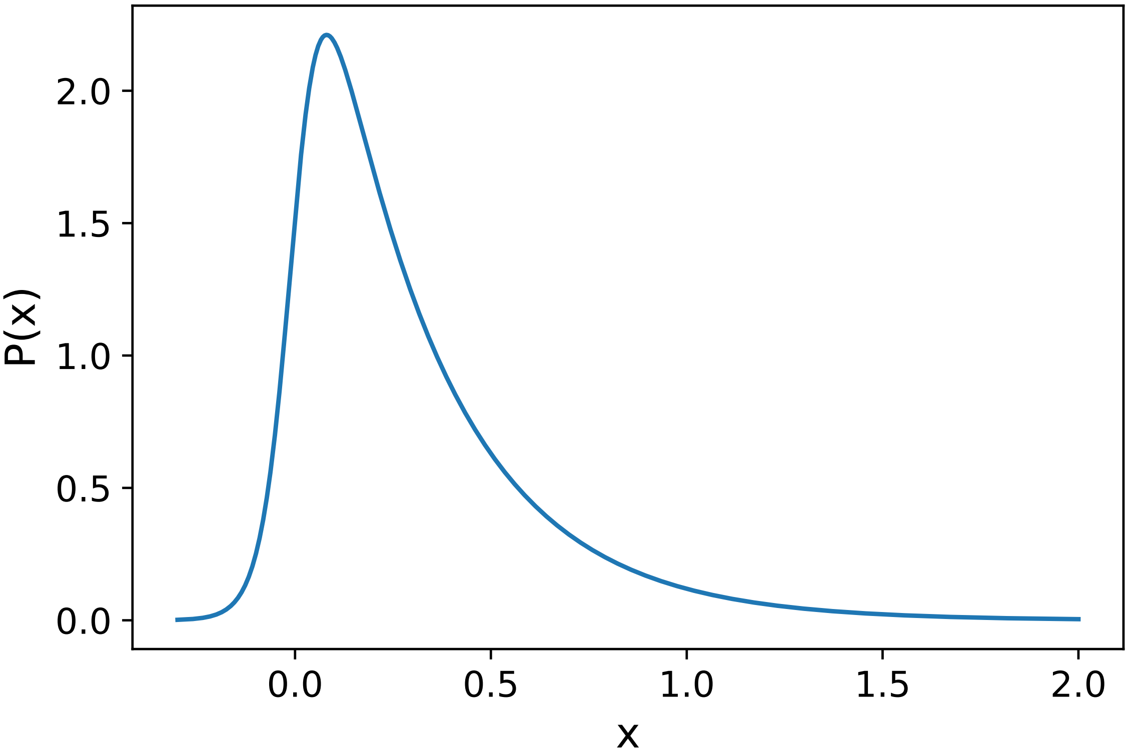

Fig. 6.1 PDF of transformed parameter x, where $A_V = \text{Softplus}(x)$

methods. This custom distribution could then be incorporated into the Pyro model in the place of sampling A_V from an exponential prior, and the model was fit as usual.

After sampling a value of x, the value of A_V is determined from x using the softplus function, and then the BayeSN model generates a light cure using this value of A_V and other fit parameters as usual. Instead of returning a posterior distribution in A_V, the model will return a Gaussian posterior of x, from which samples can be transformed to A_V using the softplus function. A point estimate of A_V was determined by taking 1,000 samples of x from the posterior distribution, applying the softplus transformation, and then taking the mean value of A_V from these transformed samples.

To evaluate the effect of this new transformation, light curves for a supernova with a low value of $A_V = 0.022$ were generated. This light curve was fitted with three methods; MCMC with an exponential prior, the original VI model with the log transform, and VI with the softplus transform.

Figure 6.2 shows the resulting posterior distributions of A_V using each of the three methods, compared with the true values from the simulation. The original VI model with the log transformation has virtually no probability density less than the true value of A_V, while using the softplus transformation leads to an approximate posterior distribution much more similar to that generated with MCMC.

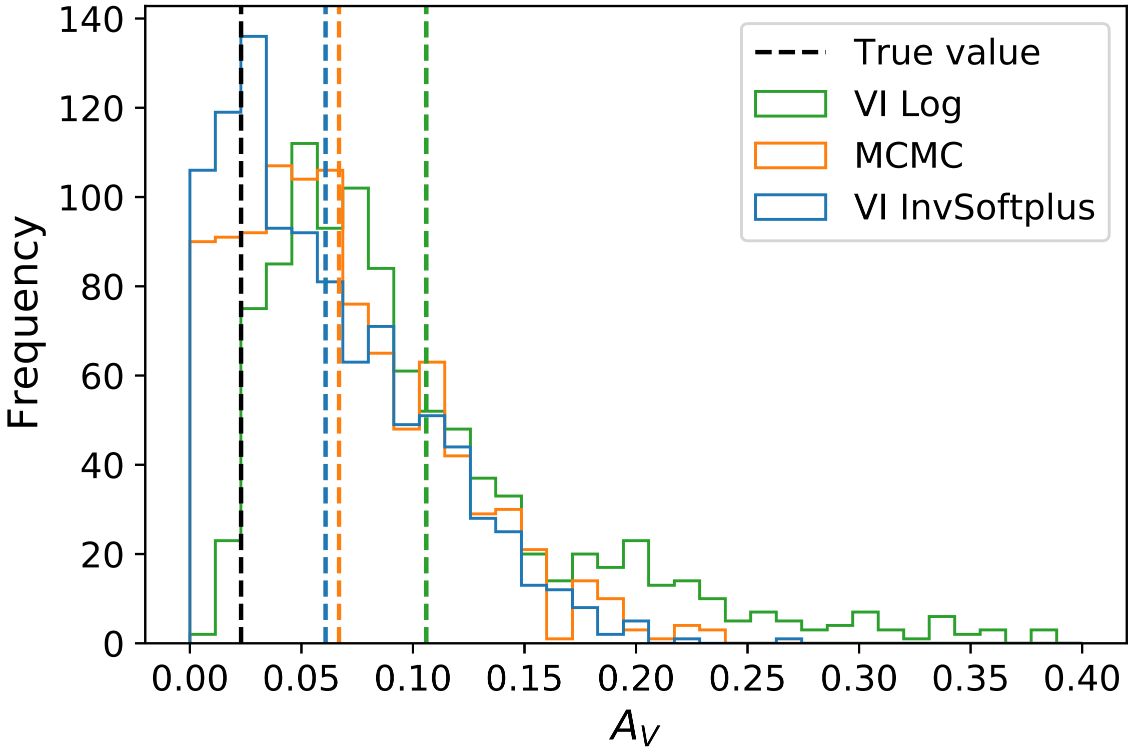

Fig. 6.2 Histogram of 1,000 A_V samples for a simulated supernova fitted with MCMC with an exponential prior, VI with a log transformation, and VI with a softplus transformation. The simulated supernova had a true value of $A_V = 0.022$, denoted with the dashed black line. Dashed colored lines refer to the mean value from each corresponding distribution.

To further assess the effect of the softplus tranform, we simulate 100 supernovae and fit them with a modified VI model incorporating this transform. Figure 6.3 shows a comparison between the “true" value of x and the mean value of x determined from the VI model using the softplus transformation. While the values seem to largely agree, there is a noticeable trend of the fit values being slightly larger than the true values of x, and there are three noticeable points with larger deviations from the true values.

This effect is then propagated to the estimates of A_V, which can be seen in the rightmost column of Figure 6.4. Many of the same biases in A_V as well as the negative bias in θ are seen in Figure 4.3 are reflected in these results, suggesting that, given the way we are currently calculating point estimates, the softplus transform may not provide the best mitigation for A_V bias.

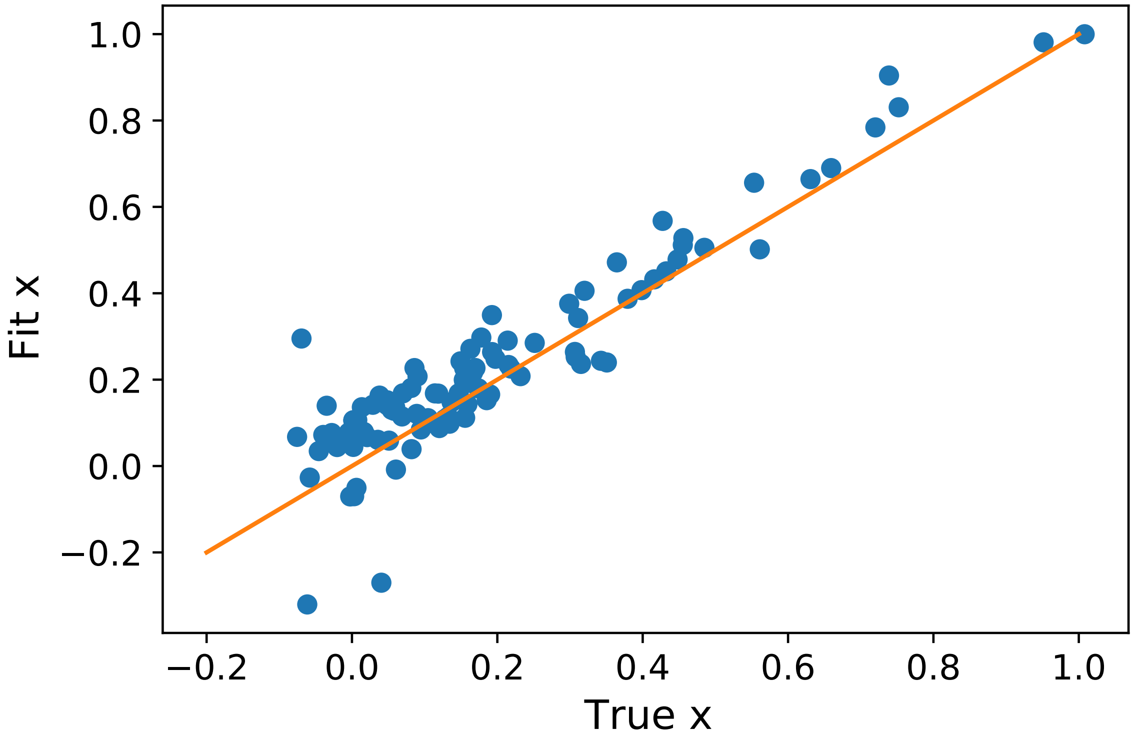

Fig. 6.3 Fit vs. true values of x, which is transformed via the softplus function to A_V for 100 simulated supernovae. Orange line denotes $y = x$ for comparison.

While comparing the model fit to the true value from simulations is a valuable analysis, in reality we can only hope to match the estimates from MCMC for real supernovae. Figures 6.5 and 6.6 compare the A_V determined from the VI with the softplus transformation to the A_V estimates determined via MCMC for the Foundation dataset. There is a slight negative bias among the entire dataset; however, low values of A_V seem to be more equally over- and underestimated, suggesting general improvement in biases associated with the exponential prior on A_V.

6.3 Asymmetric Laplace Distribution

6.3.1 Theoretical overview

The Laplace distribution is known as the "double exponential" distribution, essentially putting two exponential functions back-to-back along the vertical axis. Its PDF takes the form:

$$p(A_V) = \frac{1}{2\tau}\exp(-|A_V|/\tau) \tag{6.6}$$

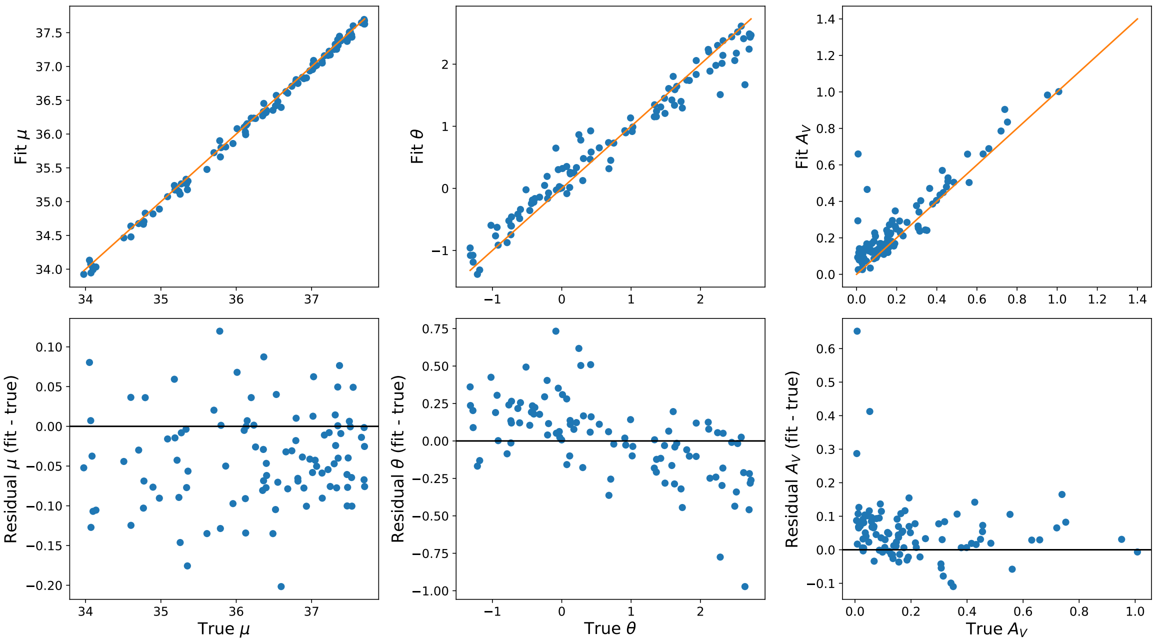

Fig. 6.4 Comparison of true vs. fit values of μ, θ, and A_V for 100 simulated supernovae using VI with a softplus transformation on A_V. Top row plots true vs. fit values with a line at $y = x$ for comparison, bottom row plots residuals (fit - true) as a function of the true parameter values.

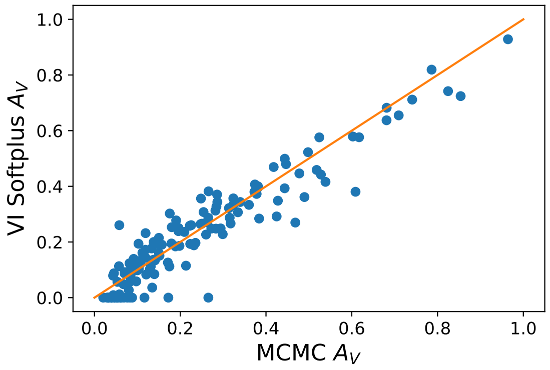

Fig. 6.5 Comparison of MCMC A_V and A_V determined from the VI Softplus model for the Foundation dataset (Foley et al., 2018). $y = x$ is plotted for comparison.

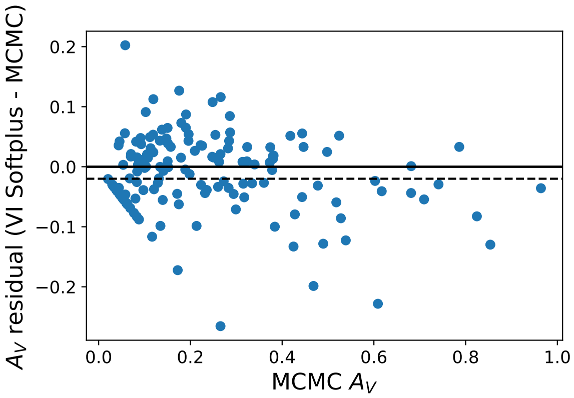

Fig. 6.6 A_V Residuals (VI Softplus - MCMC) vs MCMC A_V for the VI Softplus model on the Foundataion dataset. Dotted line denotes the median residual value.

where τ is the exponential scale parameter.

An asymmetric Laplace distribution simply puts two different exponential distributions back-to-back, each with its own scale parameter as described above. Thus, the PDF becomes:

$$p(A_V) = \frac{1}{\tau'(\kappa + 1/\kappa)} \begin{cases} -\exp(A_V/\kappa\tau') & A_V < 0 \\ \exp(\kappa A_V/\tau') & A_V > 0 \end{cases} \tag{6.7}$$

where the new scale τ' is the geometric mean of the left and right scales and the asymmetry parameter κ denotes the square of the ratio of the two scales. The two exponential distributions meet exactly at zero to ensure that the distribution is continuous.

As previously mentioned, the tendency to overestimate of A_V values close to zero comes from approximating the posterior with a Gaussian distribution and taking the mean as a point estimate. Here, we use the asymmetric Laplace distribution to create a “tail" that includes some probability at $x < 0$ in to circumvent the transform usually required to maintain $A_V > 0$. By doing this, and having A_V defined over both positive and negative values, we avoid doing a transform to impose a positive constraint as we have in the past. Note that the true values of A_V are always positive; the inclusion of the additional distribution on the negative side is inherently unphysical, but serves to reduce bias in estimates for small true A_V values.

Figure 6.7 shows the PDF for an exponential distribution with scale $\tau = 0.194$ (as in Thorp et al. (2021)) and for an asymmetric Laplace distribution with a mean scale value of $\tau/4 = 0.0485$ and an asymmetry parameter of 1/4. The values of the parameters in the asymmetric Laplace distribution were chosen to replicate the exponential PDF closely for $x > 0$, while including a small amount of probability for $x < 0$. As seen in Figure 6.7, the asymmetric Laplace distribution does not go directly to zero at $x = 0$ like the exponential distribution, but instead has a much steeper negative exponential distribution that provides slightly a more gradual decline of the PDF.

6.3.2 Implementation

The existing Pyro model to fit a single supernova with VI discussed in Chapters 4 and 5 was modified to sample A_V from an Asymmetric Laplace prior with τ' = $\tau/4$ and $\kappa = 1/4$ instead of the previously used exponential prior.

As in the previous section, a low-A_V supernova was generated to investigate the effect of the asymmetric Laplace prior compared to previous models. Figure 6.8 shows the posterior A_V distributions for the same simulated supernova as described in Section 6.2.2 fit with the original VI model (with an exponential prior), the new VI model (with an asymmetric Laplace prior), and with MCMC with the Exponential prior for comparison. While all three approaches overestimate the value of A_V in their point estimates (denoted by colored dashed

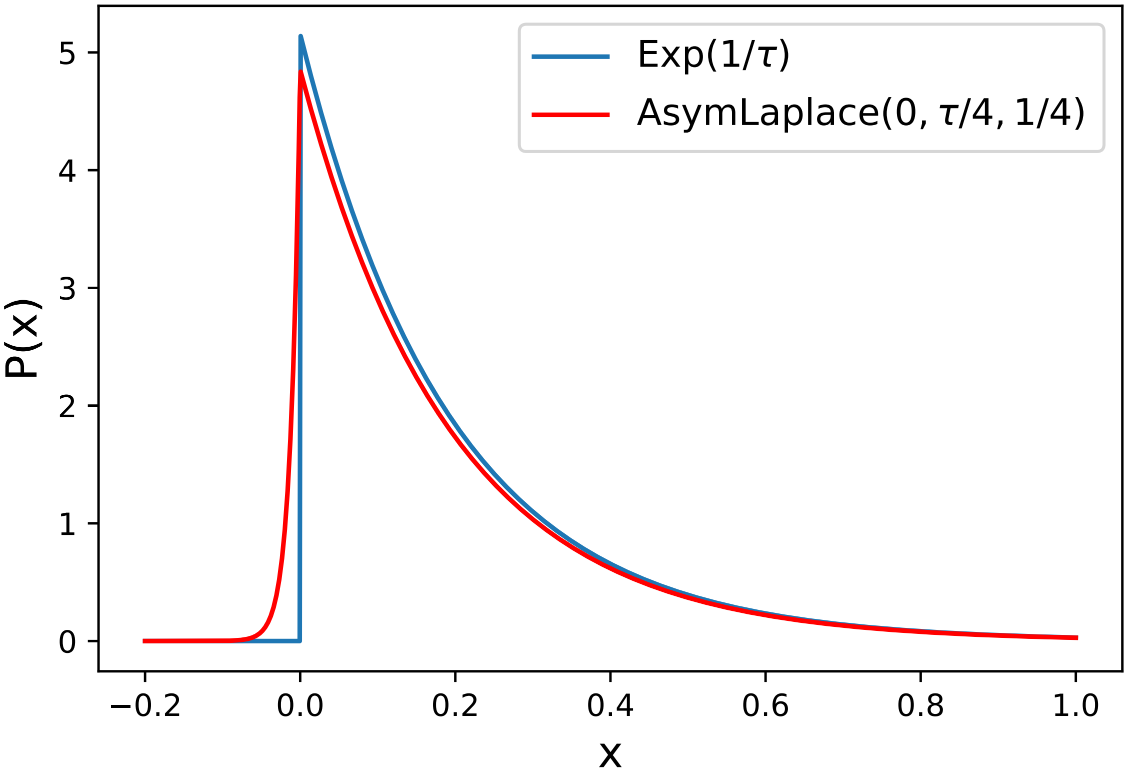

Fig. 6.7 PDF of Exponential distribution with scale $\tau = 0.194$ (as in Thorp et al. (2021)) and Asymmetric Laplace distribution with mean scale $\tau/4 = 0.0485$ and asymmetric parameter (ratio of left to right sales) of 1/4.

lines), the model using the asymmetric Laplace prior provides the desired truncated Gaussian shape when $A_V > 0$ while the model with the exponential prior has virtually no probability density lower than the true A_V value. This suggests that the inclusion of probability when $A_V < 0$ could produce more accurate estimates of for low true values of A_V, at the cost of the distribution having an unphysical “tail".

To further assess the effect of the asymmetric Laplace prior on A_V, a simulated population of 100 supernovae was created using the same procedure as described in Chapter 4. This simulated sample was then fitted one at a time using the modified VI model with an asymmetric Laplace prior, as described above. The resulting fits for these simulations can be seen in Figure 6.9, and generally show a slightly improved bias in A_V, with a relatively equal distribution of positive and negative residuals. The sloping effect in the negative residuals comes from the fact that A_V is constrained to be positive and thus is less likely to be underestimated for true values closer to zero. The distributions of θ and μ seem to be mainly unaffected compared to Figure 4.3, suggesting that use of the asymmetric Laplace

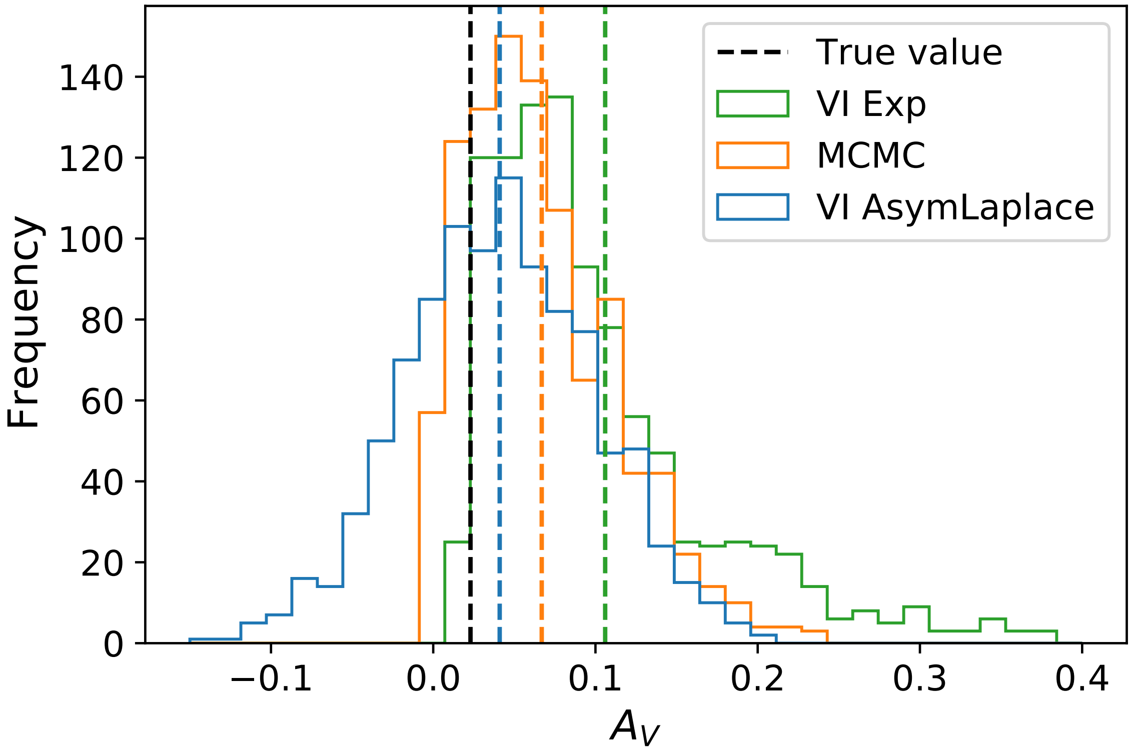

Fig. 6.8 Histogram of 1,000 A_V samples for a simulated supernova fitted with MCMC with an exponential prior, VI with an exponential prior, and VI with an asymmetric Laplace prior. The simulated supernova had a true value of $A_V = 0.022$, denoted with the dashed black line. Dashed colored lines refer to the mean value from each corresponding distribution.

distribution to model A_V could improve overall model performance. To impose the physical constraint that $A_V > 0$ after inference, a rejection sampling approach could be used to only take the positive samples from the resulting approximate posterior.

As before, we also compare the Asymmetric Laplace VI model with the MCMC fit for the Foundation dataset. Figures 6.10 and 6.11 show comparisons and residuals for the A_V estimates for MCMC with an exponential prior and VI with an Asymmetric Laplace prior. There is a more significant negative bias than in Figure 6.6, suggesting that the addition of probability when $A_V < 0$ could be causing the point estimates to be lower than they would be otherwise.

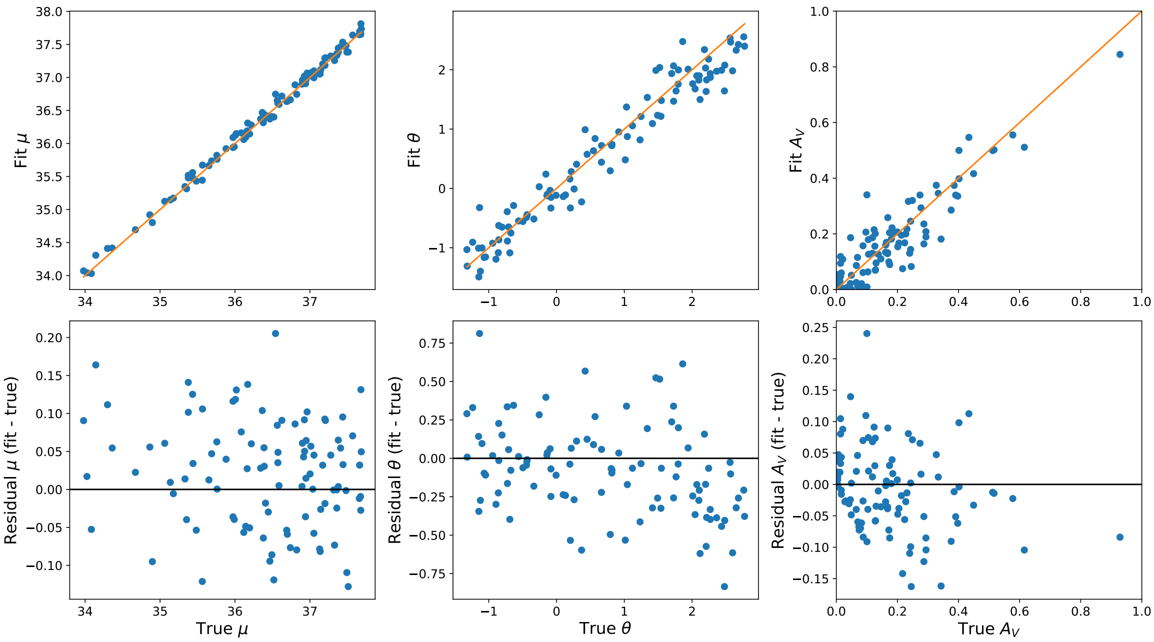

Fig. 6.9 Comparison of true vs. fit values of μ, θ, and A_V for 100 simulated supernovae using VI with an asymmetric Laplace prior on A_V. Top row plots true vs. fit values with a line at $y = x$ for comparison, bottom row plots residuals (fit - true) as a function of the true parameter values.

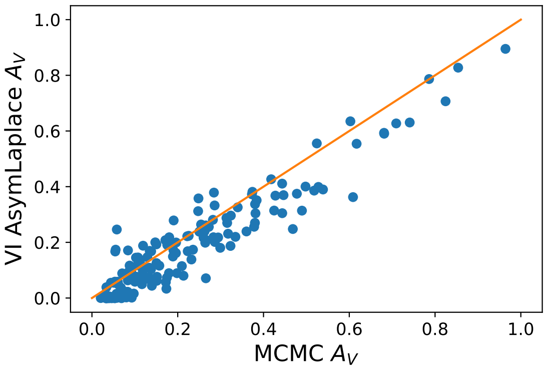

Fig. 6.10 Comparison of MCMC A_V and A_V determined from the VI Asymmetric Laplace model for the Foundation dataset (Foley et al., 2018). $y = x$ is plotted for comparison.

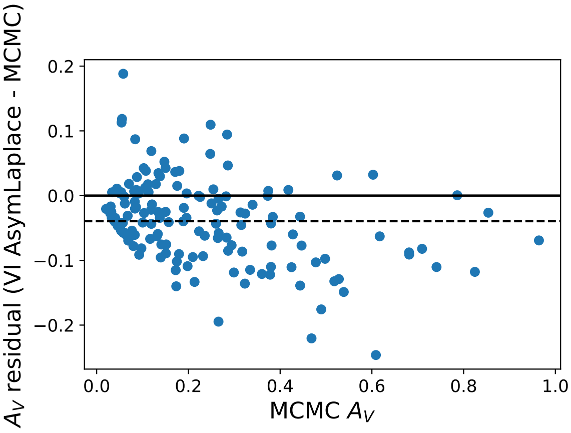

Fig. 6.11 A_V Residuals (VI Asymmetric Laplace - MCMC) vs MCMC A_V for the VI Asymmetric Laplace model on the Foundation dataset. Dotted line denotes the median residual value.

Chapter 7

Conclusion

In this work, we develop a VI model to fit probabilistic spectrotemporal models to Type Ia supernova time-series data and determining posterior distributions over the distance parameter μ, the shape parameter θ, and the dust extinction parameter A_V. We implement the BayeSN model outlined in Mandel et al. (2022) and Thorp et al. (2021) to create a probabilistic generative model of the light curves. Using Pyro's Stochastic Variational Inference functionality, we find the approximate posterior of the model parameters conditional on the observed data. We apply and validate our VI implementation on both simulated and real Type Ia supernova datasets, and explore in detail the trends and biases of the model results compared to ground truth values and MCMC sampled posteriors.

We use the VI model to fit the Foundation dataset consisting of 157 low-redshift supernovae (Foley et al., 2018; Jones et al., 2019). Fitting this dataset alone showed that the VI model had generally good agreement with MCMC results across all parameters, with a Hubble diagram scatter identical to that of previous MCMC analyses (Thorp et al., 2021). This indicates that VI is a promising approach for fitting supernova light curves and can be used for precision cosmology. Future surveys, such as the Young Supernova Experiment, will collect even more data of the same telescope and observing strategy as the Foundation dataset, with hopes of advancing understanding of dark energy and estimating its associated quantities (Jones et al., 2021). Thus, the model's performance on the Foundation dataset is promising for its future application to novel datasets for precise estimates of cosmological quantities.

In analyses on both simulated data (Chapter 4) and real supernova data (Chapter 5), the model generally fit the data well, with a multivariate Gaussian surrogate posterior providing a reasonable approximation of the true posterior. In both cases, there is a slight tendency to overestimate low A_V values that likely propagates to biases in the other fit parameters, including the distance μ. Much of the bias is likely caused by approximating the posterior of a positive variable as a Gaussian and then using the posterior mean as the point estimate. To constrain A_V to be positive, a log transform is applied to an unconstrained variable, leading to overestimation in cases where the true value of A_V is close to zero. In the observed data, selection effects propagate this bias from A_V to μ, as supernovae that are farther away are less likely to be detected if they also have high levels of dust extinction.

In Chapter 6 we discuss in detail potential modifications to the VI model that could mitigate this bias in the A_V parameter. We explore the trade-off between using a physically informed prior and one that incorporates non-physical probability in hopes of reducing bias. We first use an inverse-softplus transform to reparameterize A_V in terms of a normally-distributed variable, which yields a slightly improved bias but did not show a significant improvement in accuracy on either simulated on real datasets. We then replace the exponential prior on A_V with an Asymmetric Laplace prior that introduces some probability with $A_V < 0$. This new prior decreased overestimation at low A_V, but also included a negative bias when tested on the Foundation dataset. Overall, more analysis should be done on the source of this A_V bias and what would improve the performance of the VI model.

There are several interesting avenues for future work. In simulation-based analysis, we compare the VI fit results to the true value used in the simulation initialization, while in analyzing real supernova datasets, we compare to MCMC chains from Thorp et al. (2021). These analyses could likely be strengthened by generating MCMC fits for simulated supernovae and comparing the MCMC posterior samples to the true values to understand the biases and limitations of MCMC. These additional calibration tests would allow for a more robust understanding of the performance of the VI models across different datasets.

Generally, in this work, we have compared point estimates of the fit parameters between different models, taking the mean of each distribution. In reality, these estimators may be biased, and it would be better to compare the entire VI posterior distribution against the MCMC distribution for each parameter. This could be done by calculating the KL Divergence between the VI posterior and MCMC distributions to evaluate the similarity between the two distributions. Since MCMC just returns a number of samples, this would require using a Kernel Density Estimator to extrapolate a continuous PDF over the samples. This comparison of the entire distributions could yield even better insights about the performance of the VI model.

In Chapters 4 and 5, we outline work done on both simulated and real datasets of Type Ia supernovae using a VI model to fit individual supernova-level parameters, using previously trained values of hyperparameters such as R_V and τ_A. In reality, these population-level parameters also need to be optimized by fitting the joint distribution over all population- and individual-level parameters.

Figure 7.1 shows a probabilistic graphical model of the hierarchical model to infer the population-level parameters. Instead of assuming a single R_V value for all supernovae, individual values of R_V are drawn from a distribution with a population-level mean and standard deviation. Additionally, we are fitting the population value for τ, the scale of the exponential distribution of A_V, where we have previously just assumed the value from Thorp et al. (2021).

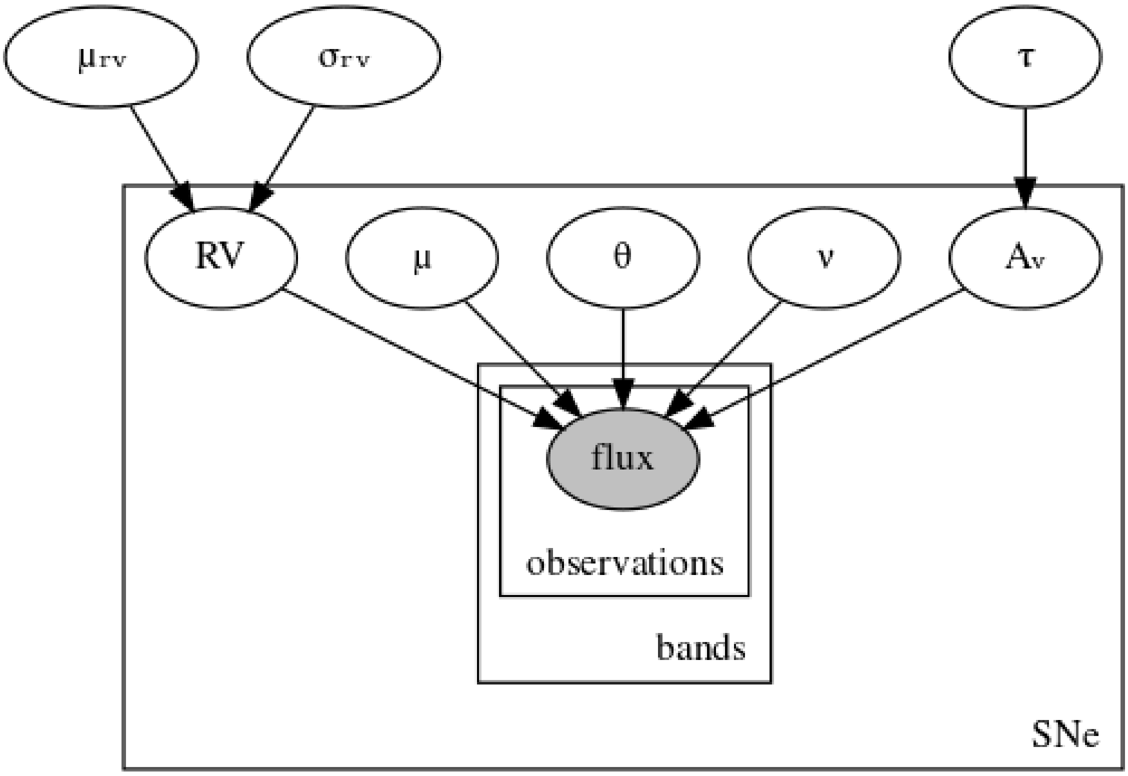

Fig. 7.1 PGM of the population-level VI model for BayeSN

This hyperparameter training would be a significant step forward from the single-supernova model used in this work; to fit the entire Foundation dataset would require training of 4,399 parameters. Yet, it can likely be achieved using the same building blocks that the smaller model is made of, utilizing Pyro's model building and VI functionality. Larger-scale problems like this are where VI could demonstrate an even more significant improvement in runtime compared to conventional approaches.

Following the work of Boone (2021), parts of the generative model from Thorp et al. (2021) used in this work could be replaced with a variational autoencoder. Instead of using the pre-determined hierarchical Bayesian model to express a supernova light curves in terms of a few latent latent variables, a decoder neural network could be used to determine the best latent representation for a light curve. This latent representation would likely would a have physical interpretation, although it may require reparameterization to be expressed in terms of commonly known physical parameters (see Boone (2021), Section 4). An encoder network could then be used to reconstruct a fitted light curve based on its latent representation. While the hyperparameters of the VAE would need to be optimized much like those in our VI model, choosing an appropriate network architecture would be an additional challenge in this approach.

Overall, this work demonstrates the potential for VI to improve model efficiency and performance for fitting type Ia supernova light curves. In the future, it can continue to be optimized and applied to other open questions in precision cosmology.

References

Agarap A. F., 2018, Deep Learning using Rectified Linear Units (ReLU), doi:10.48550/ARXIV.1803.08375, https://arxiv.org/abs/1803.08375

Astropy Collaboration et al., 2013, A&A, 558, A33

Astropy Collaboration et al., 2018, AJ, 156, 123

Avelino A., Friedman A. S., Mandel K. S., Jones D. O., Challis P. J., Kirshner R. P., 2019, ApJ, 887, 106

Bingham E., et al., 2018, arXiv e-prints, p. arXiv:1810.09538

Bishop C., 2016, Pattern Recognition and Machine Learning. Information Science and Statistics, Springer New York, https://books.google.co.uk/books?id=kOXDtAEACAAJ

Blei D. M., Kucukelbir A., McAuliffe J. D., 2016, arXiv e-prints, p. arXiv:1601.00670

Boone K., 2021, AJ, 162, 275

Breivik K., et al., 2022, arXiv e-prints, p. arXiv:2208.02781

Burns C. R., et al., 2018, ApJ, 869, 56

Cardelli J. A., Clayton G. C., Mathis J. S., 1989, ApJ, 345, 245

Dhawan S., Jha S. W., Leibundgut B., 2018, A&A, 609, A72

Dillon J. V., et al., 2017, TensorFlow Distributions, doi:10.48550/ARXIV.1711.10604, https://arxiv.org/abs/1711.10604

Draine B. T., 2003, ARA&A, 41, 241

Dvorkin C., et al., 2022, arXiv e-prints, p. arXiv:2203.08056

Feeney S. M., Mortlock D. J., Dalmasso N., 2018, MNRAS, 476, 3861

Fitzpatrick E. L., 1999, PASP, 111, 63

Foley R. J., et al., 2018, MNRAS, 475, 193

Foreman-Mackey D., 2016, The Journal of Open Source Software, 1, 24

Foreman-Mackey D., Hogg D. W., Lang D., Goodman J., 2013, PASP, 125, 306

Freedman W. L., 2021, ApJ, 919, 16

Freedman W. L., et al., 2001, ApJ, 553, 47

Ge H., Xu K., Ghahramani Z., 2018, in International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain. pp 1682–1690, http://proceedings.mlr.press/v84/ge18b.html

Gelman A., Carlin J. B., Stern H. S., Dunson D. B., Vehtari A., Rubin D. B., 2014, Bayesian Data Analysis, 3rd edn. Chapman & Hall/CRC Texts in Statistical Science, CRC Press/Taylor & Francis

Hafner D., Lillicrap T., Ba J., Norouzi M., 2019, arXiv e-prints, p. arXiv:1912.01603

Ho J., Jain A., Abbeel P., 2020, arXiv e-prints, p. arXiv:2006.11239

Hogg D. W., 1999, arXiv e-prints, pp astro–ph/9905116

Hogg D. W., 2022, arXiv e-prints, p. arXiv:2206.00989

Hsiao E. Y., 2009, PhD thesis, Univ. Victoria

Hsiao E. Y., Conley A., Howell D. A., Sullivan M., Pritchet C. J., Carlberg R. G., Nugent P. E., Phillips M. M., 2007, ApJ, 663, 1187

Hubble E., 1929, Proceedings of the National Academy of Science, 15, 168

Huber S., et al., 2022, A&A, 658, A157

Ivezić Ž., et al., 2019, ApJ, 873, 111

Jha S. W., Maguire K., Sullivan M., 2019, Nature Astronomy, 3, 706

Jones D. O., et al., 2019, ApJ, 881, 19

Jones D. O., et al., 2021, ApJ, 908, 143

Jordan M. I., Ghahramani Z., Jaakkola T. S., Saul L. K., 1999, Machine learning, 37, 183

Kasen D., Röpke F. K., Woosley S. E., 2009, Nature, 460, 869

Kim A. G., et al., 2013, ApJ, 766, 84

Kingma D. P., Ba J., 2014, Adam: A Method for Stochastic Optimization, doi:10.48550/ARXIV.1412.6980, https://arxiv.org/abs/1412.6980

Kingma D. P., Welling M., 2014, CoRR, abs/1312.6114

Knox L., Millea M., 2020, Phys. Rev. D, 101, 043533

Kullback S., Leibler R. A., 1951, The annals of mathematical statistics, 22, 79

Larsen A. B. L., Sønderby S. K., Larochelle H., Winther O., 2016, in Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48. ICML'16. JMLR.org, p. 1558–1566

Lemaître G., 1927, Annales de la Société Scientifique de Bruxelles, 47, 49

Liu R., McAuliffe J. D., Regier J., 2021, arXiv e-prints, p. arXiv:2102.02409

Mandel K. S., Wood-Vasey W. M., Friedman A. S., Kirshner R. P., 2009, ApJ, 704, 629

Mandel K. S., Narayan G., Kirshner R. P., 2011, ApJ, 731, 120

Mandel K. S., Thorp S., Narayan G., Friedman A. S., Avelino A., 2022, MNRAS, 510, 3939

Paisley J., Blei D. M., Jordan M. I., 2012, in Proceedings of the 29th International Coference on International Conference on Machine Learning. ICML'12. Omnipress, Madison, WI, USA, p. 1363–1370

Paszke A., et al., 2019, in Wallach H., Larochelle H., Beygelzimer A., d'Alché-Buc F., Fox E., Garnett R., eds, , Advances in Neural Information Processing Systems 32. Curran Associates, Inc., pp 8024–8035, http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Perlmutter S., et al., 1999, The Astrophysical Journal, 517, 565–586

Phillips M. M., 1993, ApJ, 413, L105

Planck Collaboration et al., 2020, A&A, 641, A6

Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 2007, in Numerical recipes: the art of scientific computing, 3rd Edition.

Regier J., Miller A. C., Schlegel D., Adams R. P., McAuliffe J. D., Prabhat 2018, arXiv e-prints, p. arXiv:1803.00113

Riess A. G., et al., 1998, The Astronomical Journal, 116, 1009–1038

Riess A. G., et al., 2016, ApJ, 826, 56

Riess A. G., et al., 2022, ApJ, 934, L7

Rizzato M., Sellentin E., 2022, arXiv e-prints, p. arXiv:2203.05009

Rose B. M., et al., 2021, arXiv e-prints, p. arXiv:2111.03081

Sanchez-Lengeling B., Aspuru-Guzik A., 2018, Science, 361, 360

Schlafly E. F., Finkbeiner D. P., 2011, ApJ, 737, 103

Schlafly E. F., et al., 2016, ApJ, 821, 78

Schlegel D. J., Finkbeiner D. P., Davis M., 1998, ApJ, 500, 525

Spiegelhalter D. J., Thomas A., Best N. G., Gilks W. R., 1995, in BUGS - Bayesian inference Using Gibbs Sampling Version 0.50.

Stan Development Team 2018, Stan Modeling Language Users Guide and Reference Manual, Version 2.18.0, http://mc-stan.org/

The LSST Dark Energy Science Collaboration et al., 2018, arXiv e-prints, p. arXiv:1809.01669

Thorp S., Mandel K. S., Jones D. O., Ward S. M., Narayan G., 2021, Monthly Notices of the Royal Astronomical Society, 508, 4310

Vincenzi M., Sullivan M., Firth R. E., Gutiérrez C. P., Frohmaier C., Smith M., Angus C., Nichol R. C., 2019, MNRAS, 489, 5802

Wainwright M. J., Jordan M. I., et al., 2008, Foundations and Trends® in Machine Learning, 1, 1

Appendix A

Derivation of Laplace Approximation

The Laplace Approximation approximates a likelihood function as Gaussian centered around the maximum likelihood estimate. For some log-likelihood function ℓ, with parameters θ, we can perform a second-order Taylor expansion about the MLE estimate $\hat{\theta}$:

$$\ell(\theta) \approx \ell(\hat{\theta}) + \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta}) + \frac{1}{2} \left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 + ... \tag{A.1}$$

By the definition of MLE, we know that

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0 \tag{A.2}$$

Thus, the Laplace Approximation of the log-likelihood function is:

$$\ell(\theta) \approx \ell(\hat{\theta}) + \frac{1}{2} \left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta=\hat{\theta}} (\theta - \hat{\theta})^2 \tag{A.3}$$

The likelihood function is thus transformed to a Gaussian distribution as follows:

$$\mathscr{L}(\theta) \approx \mathscr{L}(\hat{\theta}) \exp\left(\frac{1}{2}(\theta - \hat{\theta})^2 \left(\frac{\partial^2 \ell}{\partial \theta^2}\right)\right) \approx \mathrm{N}(\hat{\theta}, \hat{\sigma}^2) \tag{A.4}$$

which can be expressed with the MLE as mean and an arbitrary variance that can be solved for (Gelman et al., 2014).

Appendix B

Distance modulus relation

The total makeup of the universe can be expressed as a sum of different densities:

$$\Omega_M + \Omega_\Lambda + \Omega_k = 1 \tag{B.1}$$

Where Ω_M is the density of matter, Ω_Λ is the density of dark energy, and Ω_k is the curvature of the universe. The sign of Ω_k in particular has important implications for the shape and concavity of the universe, and how other cosmological quantities interact with each other (Hogg, 1999).

The dimensionless comoving distance denotes the distance to an astronomical object relative to the expansion of the universe. It is determined from the object's redshift z as follows:

$$\tilde{d}(z;,\Omega_M,\Omega_\Lambda,w) = \int_0^z \frac{dz'}{\sqrt{\Omega_M(1+z')^3 + \Omega_k(1+z')^2 + \Omega_\Lambda(1+z')^{3(w+1)}}} \tag{B.2}$$

where w is the dark energy equation-of-state parameter (Hogg, 1999).

The luminosity distance d_L denotes the distance to an object based on its measured luminosity, or brightness. It is derived from the comoving distance as follows:

$$\tilde{d}_L(z_s;\Omega_M,\Omega_\Lambda,w) = \begin{cases} |\Omega_k|^{-\frac{1}{2}} \sinh\left[\sqrt{|\Omega_k|}\tilde{d}(z;,\Omega_M,\Omega_\Lambda,w)\right] & \Omega_k < 0 \\ \tilde{d}(z;,\Omega_M,\Omega_\Lambda,w) & \Omega_k = 0 \\ |\Omega_k|^{-\frac{1}{2}} \sin\left[\sqrt{|\Omega_k|}\tilde{d}(z;,\Omega_M,\Omega_\Lambda,w)\right] & \Omega_k > 0 \end{cases} \tag{B.3}$$

The distance modulus μ can then be obtained using Equation 2.6 as described previously:

$$\mu = 25 + 5\log_{10}\left[\frac{c}{H_0}\tilde{d}_L(z_s;\Omega_M,\Omega_\Lambda,w)\right] \mathrm{Mpc}^{-1} \tag{B.4}$$

Appendix C

Code

All original code written for this project is available in the following GitHub repository:

```
https://github.com/asmuzsoy/variational_bayesn
```

Note that the file `spline_hsiao_fns.py` contains functions written by Stephen Thorp.