# Variational Inference for Acceleration of SN Ia Photometric Distance Estimation with BayeSN

Ana Sofía M. Uzsoy,[1,2,3]⋆ Stephen Thorp,[4,1] Matthew Grayling,[1] and Kaisey S. Mandel[1,5]

[1]*Institute of Astronomy and Kavli Institute for Cosmology, Madingley Road, Cambridge, CB3 0HA, UK*
[2]*Department of Engineering, University of Cambridge, Trumpington Street, Cambridge, CB2 1PZ, UK*
[3]*Center for Astrophysics | Harvard & Smithsonian, 60 Garden Street, Cambridge, MA 02138, USA*
[4]*The Oskar Klein Centre, Department of Physics, Stockholm University, AlbaNova University Centre, SE 106 91 Stockholm, Sweden*
[5]*Statistical Laboratory, DPMMS, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK*

ABSTRACT

Type Ia supernovae (SNe Ia) are standarizable candles whose observed light curves can be used to infer their distances, which can in turn be used in cosmological analyses. As the quantity of observed SNe Ia grows with current and upcoming surveys, increasingly scalable analyses are necessary to take full advantage of these new datasets for precise estimation of cosmological parameters. Bayesian inference methods enable fitting SN Ia light curves with robust uncertainty quantification, but traditional posterior sampling using Markov Chain Monte Carlo (MCMC) is computationally expensive. We present an implementation of variational inference (VI) to accelerate the fitting of SN Ia light curves using the BayeSN hierarchical Bayesian model for time-varying SN Ia spectral energy distributions (SEDs). We demonstrate and evaluate its performance on both simulated light curves and data from the Foundation Supernova Survey with two different forms of surrogate posterior – a multivariate normal and a custom multivariate zero-lower-truncated normal distribution – and compare them with the Laplace Approximation and full MCMC analysis. To validate of our variational approximation, we calculate the pareto-smoothed importance sampling (PSIS) diagnostic, and perform variational simulation-based calibration (VSBC). The VI approximation achieves similar results to MCMC but with an order-of-magnitude speedup for the inference of the photometric distance moduli. Overall, we show that VI is a promising method for scalable parameter inference that enables analysis of larger datasets for precision cosmology.

Key words: supernovae: general – distance scale – dust, extinction – methods: statistical

1 INTRODUCTION

1.1 Scientific Context

Type Ia supernovae (SNe Ia) are well known as "standardizable candles" – astronomical objects for which there are strong correlations between their absolute brightness and apparent features of their light curves (e.g. Rust 1975; Pskovskii 1977; Phillips 1993; Tripp 1998). They are widely used in precision cosmology, and played a key role in the discovery of dark energy (Riess et al. 1998; Perlmutter et al. 1999). In the current debate (for a review see e.g. Knox & Millea 2020; Freedman 2021) over the value of the Hubble–Lemaître constant H_0, the SN Ia distance ladder anchored by Cepheid variables (Riess et al. 2022) is in tension with the cosmic microwave background (Planck Collaboration et al. 2020).

Over the past few decades, higher fidelity and greater volumes of data have been obtained for SNe Ia (e.g. Hamuy et al. 1996a; Riess et al. 1999; Astier et al. 2006; Frieman et al. 2008; Kessler et al. 2009b; Hicken et al. 2012; Friedman et al. 2015; Scolnic et al. 2018; Foley et al. 2018; Brout et al. 2019; Abbott et al. 2024). This has allowed the sophistication of SN Ia models to grow from simple scaling relations and templates (e.g. Rust 1975; Pskovskii

1977; Phillips 1993; Hamuy et al. 1996b), to learned representations of light curves (e.g. Riess et al. 1995, 1996a; Jha et al. 2007; Mandel et al. 2009, 2011; Burns et al. 2011) and spectral energy distributions (SEDs; e.g. Guy et al. 2005, 2007; Conley et al. 2008; Saunders et al. 2018; Kenworthy et al. 2021; Boone 2021; Thorp et al. 2021; Mandel et al. 2022; Grayling et al. 2024).

For as long as SNe Ia have been used as standardizable candles for cosmology, the impact of dust in their host galaxies has been discussed extensively by astronomers (e.g. Branch & Tammann 1992; Sandage & Tammann 1993; Riess et al. 1996a,b; Phillips et al. 1999; Krisciunas et al. 2000, 2007; Mandel et al. 2011; Burns et al. 2014; Mandel et al. 2017; Brout & Scolnic 2021; Popovic et al. 2021; Thorp & Mandel 2022; Karchev et al. 2023, 2024; Grayling et al. 2024; Thorp et al. 2024). Dust along the line of sight to a supernova or other astrophysical source acts to dim the light from the source. Preferential dimming of bluer wavelengths also causes reddening of the light (see e.g. Draine 2003; Salim & Narayanan 2020, for general reviews on astrophysical dust). Dust in the Milky-Way galaxy is extensively mapped (e.g. Schlegel et al. 1998; Schlafly & Finkbeiner 2011; Schlafly et al. 2016), so it is possible to correct for its effects. However, the host galaxies of individual SNe Ia are not mapped in this way, so extinction due to host galaxy dust must be modelled probabilistically. This is one of the primary motivations for the BAYESN hierarchical SED model for SNe Ia (Thorp et al. 2021; Mandel et al.

⋆ E-mail: ana_sofia.uzsoy@cfa.harvard.edu

2022; Grayling et al. 2024). Current implementations of the BAYESN model use the No U-Turn Sampler (NUTS; Hoffman & Gelman 2014) to perform full Bayesian inference given a set of SN Ia data. This has been applied to studies of host galaxy dust (Thorp et al. 2021; Thorp & Mandel 2022; Grayling et al. 2024), SN Ia siblings (Ward et al. 2023b), estimation of the Hubble–Lemaître constant (Dhawan et al. 2023), analyses of new near-infrared (NIR) SN Ia datasets (Jones et al. 2022; Peterson et al. 2023; Thorp et al. 2024), and lensed supernova time-delay estimation (Pierel et al. 2024). A simulation-based inference (SBI) implementation of BAYESN has also been developed recently by Karchev et al. (2023, 2024). To ensure BAYESN remains scalable in anticipation of the large datasets expected from future surveys, we present here a variational implementation of the model for accelerated Bayesian inference of distances from SN Ia light curves. This is built on top of the GPU-accelerated BAYESN code developed by Grayling et al. (2024).

1.2 The Need for Speed

The datasets used in SN Ia cosmology have already been growing substantially over recent years (see e.g. Phillips et al. 2019; Dhawan et al. 2022; Aleo et al. 2023; Abbott et al. 2024). With new, powerful telescopes and space missions observing more astronomical objects than ever before, the field of astrophysics is generally becoming more data-driven. The Vera C. Rubin Legacy Survey of Space and Time (LSST) is expected to make over 32 trillion observations in over 20 billion galaxies and stars (LSST Dark Energy Science Collaboration et al. 2018; Ivezić et al. 2019), and will be transformative for SN Ia and transient science. The *Nancy Grace Roman Space Telescope* will be similarly transformative for SN Ia cosmology at NIR wavelengths (Hounsell et al. 2018; Rose et al. 2021). Scalable analysis methods are critical to maximizing the scientific output from these projects. Moreover, the "industry standard" pipelines and validation frameworks (e.g. Kessler et al. 2009a; Kessler & Scolnic 2017; Hinton & Brout 2020) that will be used in the core SN Ia cosmology analyses for next generation surveys typically require methods that can be rapidly applied to tens to hundreds of thousands of simulated SNe for bias modeling.

Over the past two decades, parameter inference problems in astrophysics and cosmology have often been framed in a Bayesian way, and solved with Markov Chain Monte Carlo (MCMC) samplers and other related numerical methods. The development and uptake of user-friendly interfaces to general purpose sampling algorithms (e.g. Lewis & Bridle 2002; Feroz et al. 2009; Foreman-Mackey et al. 2013; Handley et al. 2015; Speagle 2020) has greatly facilitated this. While posterior exploration with MCMC remains the gold-standard for Bayesian inference, it is often limited by computational time and resources.

Variational inference (VI; see §1.3 for an introductory overview) has been shown to improve computational efficiency by orders of magnitude compared to MCMC when applied to various astrophysical problems (e.g. Regier et al. 2019; Gunapati et al. 2022; Rizzato & Sellentin 2023; Liu et al. 2023). de Soto et al. (2024) use VI for parametric light curve fitting for real-time supernova classification, and Sánchez et al. (2021) use VI amortized with a neural network to fit SN Ia light curves. Villar (2022) also present an amortized inference approach, using masked auto-regressive flows (MAFs; Papamakarios et al. 2017) to approximate the posterior distribution (see e.g. Rezende & Mohamed 2015). Alternative variational methods – particularly variational auto-encoders (VAEs) – have also been applied effectively to supernovae and other transients. Boone (2021) and Villar et al. (2021) present VAEs to learn latent representations

in the context of supernova SED fitting and light curve classification, respectively. With a VAE, the neural network weights of the encoder and decoder are optimized to learn a latent representation by minimizing the average reconstruction loss across all objects in the training set. In contrast, here we use a pre-trained physically-motivated model (BayeSN) for SN Ia SEDs and light curve data and use VI to approximate the true posterior for each individual SN Ia separately.

1.3 Overview of Variational Inference (VI)

Variational Inference (VI) is a machine learning approach that approximates an intractable posterior that cannot be directly sampled with a simpler "surrogate" distribution, also known as a "guide" (Jordan et al. 1999). The parameters defining this surrogate distribution are then optimized to minimize the difference between the surrogate and true posteriors (Jordan et al. 1999). For data x, model parameters $\boldsymbol{\phi}$, and variational parameters ζ, the surrogate posterior $q_\zeta^*(\boldsymbol{\phi})$ is determined as follows:

$$q_\zeta^*(\boldsymbol{\phi}) = \underset{q_\zeta(\boldsymbol{\phi}) \in Q}{\arg\min} \; D_{\mathrm{KL}}[q_\zeta(\boldsymbol{\phi}) \, || \, p(\boldsymbol{\phi}|x)] \tag{1}$$

where $p(\boldsymbol{\phi}|x)$ is the true posterior, D_{KL} denotes the Kullback-Leibler divergence (Kullback & Leibler 1951), and Q denotes a chosen family of distributions (Blei et al. 2017). As the true posterior $p(\boldsymbol{\phi}|x)$ is intractable and thus unable to be directly computed, VI maximizes the Evidence Lower Bound (ELBO; Jordan et al. 1999; Wainwright & Jordan 2008; Blei et al. 2017) optimize the parameters ζ defining the surrogate posterior[1]:

$$\mathrm{ELBO}(q_\zeta(\boldsymbol{\phi})) = \mathbb{E}_q[\log p(\boldsymbol{\phi}, x)] - \mathbb{E}_q[\log q_\zeta(\boldsymbol{\phi})] \tag{2}$$

$$= \mathbb{E}_q[\log p(x|\boldsymbol{\phi})] - D_{\mathrm{KL}}[q_\zeta(\boldsymbol{\phi}) \, || \, p(\boldsymbol{\phi})]. \tag{3}$$

The parameters ζ (hereafter referred to as the "variational parameters") define the surrogate posterior $q_\zeta(\boldsymbol{\phi})$. In practice, these expectations over q are approximated by generating samples ("particles") of model parameters from the surrogate posterior for a given set of variational parameters and computing the Monte Carlo average over these samples. The variational parameters ζ are optimized with gradient descent using the stochastic variational inference algorithm (Hoffman et al. 2013; Wingate & Weber 2013; Kingma & Welling 2013; Ranganath et al. 2014; Kucukelbir et al. 2017). Validation of variational inference algorithms (e.g. Yao et al. 2018; Huggins et al. 2020) can be conducted using simulation-based calibration (Cook et al. 2006; Talts et al. 2018), Pareto-smoothed importance sampling (PSIS; Vehtari et al. 2024), or the Wasserstein distance (Kantorovich & Rubinstein 1958; Vaserstein 1969).

1.4 This Work

This paper investigates whether VI can be an appropriate replacement for MCMC when using the BAYESN model (Mandel et al. 2022) to estimate the distances to SNe Ia from their light curves. We use stochastic variational inference (SVI) in NUMPYRO (Bingham et al. 2019; Phan et al. 2019) to fit for two different forms of surrogate posterior: a multivariate normal distribution, and a multivariate zero-lower-truncated normal (ZLTN) distribution. We evaluate the performance of the two VI implementations, compared to MCMC and the Laplace approximation, on simulated and real data (from

---

[1] Note that in this work, log refers to natural logarithm (base e), while $\log_{10}$ specifies base 10.

the Foundation survey; Foley et al. 2018; Jones et al. 2019). We explore the trade-off between computational efficiency and model accuracy and validate the performance of VI in each case (following Yao et al. 2018). In terms of parameter estimation, both of the tested VI posterior forms perform comparably to MCMC, but with an order of magnitude speedup. We then discuss the potential implications of this new variational implementation of BAYESN and when VI should be used to approximate posterior distributions for inference of astrophysical parameters.

In Section 2, we review the details of the BAYESN model, and in Section 3 we describe our new VI implementation. In Section 4 we describe the real (§4.1) and simulated (§4.2) data. Our results are presented in Section 5, with additional discussion in Section 6, and concluding remarks in Section 7.

2 THE BAYESN MODEL

2.1 Overview

The full details of the BAYESN model can be found in Mandel et al. (2022), with subsequent model developments discussed in Thorp et al. (2021), Ward et al. (2023b) and Grayling et al. (2024). In summary, BAYESN is a hierarchical Bayesian model for the spectral energy distributions (SEDs) of Type Ia supernovae (SNe Ia). BAYESN defines a forward model for SN Ia light curves including a fixed spectral template (Hsiao et al. 2007; Hsiao 2009), a functional principal component-based representation of intrinsic SED variation, a physically-motivated parametric model for host galaxy dust extinction (Fitzpatrick 1999), and a model for residual spectro-temporal perturbations. Throughout this work, we use the version of the model trained by Thorp et al. (2021) – hereafter the T21 model – using data from the Foundation Supernova Survey (Foley et al. 2018; Jones et al. 2019).

In the BAYESN model (following Mandel et al. 2022, eq. 12), the host dust-extinguished rest-frame SED $S_s(t, \lambda_r)$, of a supernova s, is defined by

$$-2.5 \log_{10}\left(\frac{S_s(t, \lambda_r)}{S_0(t, \lambda_r)}\right) = M_0 + W_0(t, \lambda_r)$$
$$+ \delta M_s + \theta_1^s W_1(t, \lambda_r) + \epsilon_s(t, \lambda_r) + A_V^s \xi(\lambda_r; R_V^{(s)}), \quad (4)$$

with reference to a spectral template $S_0(t, \lambda_r)$ (Hsiao et al. 2007; Hsiao 2009), as a function of rest-frame phase t (relative to the time of maximum B-band brightness), and wavelength λ_r. Here, the supernova-level latent parameters are: a light curve shape parameter/functional principal component score θ_1^s; the V-band dust extinction A_V^s; a time- and wavelength-dependent residual perturbation function $\epsilon_s(t, \lambda_r)$; and a "gray" (t- and λ_r-independent) magnitude offset δM_s. BAYESN's θ_1^s parameter correlates with the slope of the SED surface and the width-luminosity relation in the optical (Mandel et al. 2022) and is closely related to the Δm_{15} parameter (i.e. the decline in magnitude in the 15 days following peak; see discussion in appendix E of Grayling et al. 2024). The population-level hyperparameters are: a warping function applied to the mean SED $W_0(t, \lambda_r)$; the first functional principal component $W_1(t, \lambda_r)$; and the Fitzpatrick (1999) dust law $\xi(A_V; R_V)$. The overall normalization factor is fixed $M_0 \equiv -19.5$. The dust law slope parameter can be assumed to be the same for all SNe (as in Mandel et al. 2022; Thorp et al. 2024), or allowed to vary on a supernova-by-supernova basis (as in Thorp et al. 2021; Thorp & Mandel 2022; Thorp et al. 2024; Grayling et al. 2024). The BAYESN model flux for a supernova s in a

passband b is given by Eq. 6 in Mandel et al. (2022):

$$f_s(t; b) = (1 + z_s) \times 10^{0.4(Z - \mu_s)}$$
$$\times \int_{\lambda_r \in b} S_s(t, \lambda_r) \times 10^{-0.4 A_{V,s}^{MW} \xi^{MW}[(1+z_s)\lambda_r]} \mathbb{B}_b[(1+z_s)\lambda_r] \lambda_r \, d\lambda_r. \quad (5)$$

Here, μ_s is the SN distance modulus, z_s is the redshift, $A_{V,s}^{MW}$ is the Milky Way dust extinction along the SN's line of sight (based on the Schlafly & Finkbeiner 2011 dust maps), ξ^{MW} is the Milky Way dust extinction law (Fitzpatrick 1999), $Z \equiv 27.5$ is a fixed zero-point (equal to that of SNANA; Kessler et al. 2009a), and $\mathbb{B}_b$ is the normalized transmission function of band b. The integral in Eq. 5 is taken over all rest frame wavelengths observable by band b.

In practice the functional parameters in the model, $W_0(t, \lambda_r)$, $W_1(t, \lambda_r)$, and $\epsilon(t, \lambda_r)$, are represented by natural cubic splines[2] defined by vectors/matrices of knots W_0, W_1, and e_s. Additionally, the time of maximum may not be perfectly known a priori, meaning a correction factor Δt_{max}^s may be needed. This gives a set of supernova-level latent parameters $\phi_s = (A_V^s, \mu_s, \delta M_s, \theta_1^s, \Delta t_{max}^s, e_s)$, and population-level hyperparameters $H = (W_0, W_1, R_V, \tau_A, \Sigma_e, \sigma_0)$. The latter three hyperparameters specify respectively the population distributions of A_V^s, e_s, and δM_s:

$$P(A_V^s | \tau_A) = \text{Exponential}(A_V^s | \tau_A), \quad (6)$$
$$P(e_s | \Sigma_e) = N(e_s | 0, \Sigma_e), \quad (7)$$
$$P(\delta M_s | \sigma_0) = N(\delta M_s | 0, \sigma_0^2). \quad (8)$$

The prior on θ_1^s is a unit Gaussian, $P(\theta_1^s) = N(\theta_1^s | 0, 1)$, and we adopt a broad and uninformative prior on distances $P(\mu_s) = N(\mu_s | \mu_{\Lambda CDM}(z_s), 5^2)$, where $\mu_{\Lambda CDM}(z_s)$ is the estimated distance under a flat ΛCDM cosmology. In this work, we adopt a Gaussian prior on the time of maximum correction[3], $P(\Delta t_{max}^s) = N(\Delta t_{max}^s | 0, 5^2)$. In this work, we are interested in BAYESN as a distance estimator (c.f. Thorp et al. 2021 §4.6; Mandel et al. 2022 §2.8) – i.e. given a supernova's photometry, what is the posterior on μ_s (and the other latent parameters in ϕ_s). For this reason, we will fix the values of the hyperparameters to their posterior mean values, $\hat{H}$, estimated by Thorp et al. (2021) from the full Foundation dataset (Foley et al. 2018; Jones et al. 2019)[4]. For a supernova s, we will have N observed fluxes $\hat{f}_s = (\hat{f}_{s,1}, \ldots, \hat{f}_{s,N})^\top$, with corresponding observation dates[5] $\hat{T}_s = (\hat{T}_{s,1}, \ldots, \hat{T}_{s,N})^\top$, measurement errors $\hat{\sigma}_s = (\hat{\sigma}_{s,1}, \ldots, \hat{\sigma}_{s,N})^\top$, and rest-frame phases $\hat{t}_s = (\hat{T}_s - \hat{T}_{ref}^s)/(1 + z_s)$. The phases are defined with reference to an estimated date of maximum light $\hat{T}_{ref}^s$. The posterior distribution of ϕ_s will be given by

$$P(\phi_s | \hat{f}_s; \hat{H}) \propto P(\phi_s | \hat{H})$$
$$\times \prod_{n=1}^N N(\hat{f}_{s,n} | f(\hat{t}_{s,n} + \Delta t_{max}^s; b_n, \phi_s, \hat{H}), \hat{\sigma}_{s,n}^2), \quad (9)$$

where

$$P(\phi_s | \hat{H}) = P(\mu_s) \times P(\theta_1^s) \times P(\Delta t_{max}^s)$$
$$\times P(A_V^s | \tau_A) \times P(e_s | \Sigma_e) \times P(\delta M_s | \sigma_0), \quad (10)$$

---

[2] The Fitzpatrick (1999) dust law is also defined as a cubic spline.
[3] This choice implies a 95% prior probability that the original maximum date was estimated to within $10 \times (1 + z_s)$ observer-frame days of the truth.
[4] These hyperparameter estimates can be found in the `BAYESN.T21` directory at `https://github.com/bayesn/bayesn-model-files`
[5] These are Modified Julian Dates (MJD) in observer-frame days.

the passband for observation n is denoted by b_n, $\Delta t^s_{\max}$ is a correction for the time of maximum[6], and the model flux $f(\hat{t}_{s,n}; b_n, \boldsymbol{\phi}_s, \hat{\boldsymbol{H}})$ is evaluated using Eq. 4 and 5. This posterior distribution is the target of the variational approximation we develop in the subsequent sections of this paper.

2.2 The Challenge of Dust Extinction

The dust extinction parameter A_V can take on any value in the range $[0,\infty)$, and can be correlated with other parameters in the posterior distribution. The exponential prior in Equation 6 is physically motivated, and favors low values[7]. For SNe Ia with little-to-no dust extinction (A_V close to zero), there can be considerable nonzero posterior density at $A_V \approx 0$. However, negative values of A_V are unphysical, and should not be permitted. This presents a challenge for variational inference, as we need an approximate posterior with the following properties:

(i) A_V constrained positive, all other parameters unbounded;
(ii) potential correlation between all parameters (i.e. "full rank"); and
(iii) non-negligible density allowed for A_V values arbitrarily close to zero.

A simple solution which satisfies points (i) and (ii) is to fit a multivariate normal to the joint posterior distribution of $\log A_V$ and all other parameters. This is one of the approximations we will consider in subsequent sections of this paper. The disadvantage of this approach is that the the approximate posterior marginal on A_V will be effectively log-normal, tending towards zero density as $A_V \to 0$. Underestimating the posterior density at $A_V = 0$ can potentially lead to overestimation of extinction in the low-dust limit. A multivariate normal across all parameters (including A_V) would satisfy points (ii) and (iii), but would allow for potentially problematic negative values of A_V. Points (i) and (iii) could be satisfied simultaneously by using a variational surrogate posterior that is a product of an independent truncated normal distribution on A_V, and a multivariate normal for all other parameters. To satisfy all three desiderata, we propose using a variational guide where a zero-lower-truncated normal (ZLTN) distribution approximates the marginal posterior on A_V, and the *conditional* distribution of the remaining parameters given A_V is multivariate Gaussian (the exact form is defined in §3.3.2). For A_V far away from zero, the ZLTN distribution recovers the same behavior as a non-truncated Gaussian distribution, but in the low-A_V limit the behavior is more desirable. Using this distribution allows for a more flexible posterior approximation, while still providing the computational speedup of VI over MCMC.

---

[6] Rather than fitting for the date of maximum $T^s_{\max}$ directly, we adopt a parameterization that is defined with reference to a prior estimate $\hat{T}^s_{\text{ref}}$. The correction factor is related to $T^s_{\max}$ by: $\Delta t^s_{\max} = (T^s_{\max} - \hat{T}^s_{\text{ref}})/(1 + z_s)$.

[7] The popular choice of an exponential prior (Eq. 6) on dust extinction was originated by Jha et al. (2007), motivated by the fact that viewing angles that produce high extinction must be rarer. Theoretical studies (Hatano et al. 1998; Commins 2004; Riello & Patat 2005) had also favored extinction values following an exponential distribution for late-type SN Ia hosts (see also Holwerda et al. 2015). This assumption has been scrutinized by several recent works (Wojtak et al. 2023; Ward et al. 2023a).

3 VARIATIONAL INFERENCE IMPLEMENTATION

In our VI implementation, we aim to fit 28 BAYESN model parameters: A_V, $\mu + \delta M$, θ_1, $\Delta t_{\max}$, and $\boldsymbol{e}$, where $\boldsymbol{e}$ is 24 entries for the residual realization. For identifiability reasons, we treat the sum of μ and δM as a single parameter, and separate them in post-processing. For each posterior form, we provide first a theoretical overview followed by the implementation details. In general, the BAYESN model was implemented using the NUMPYRO probabilistic programming language. Within NUMPYRO we use the SVI class to perform Stochastic Variational Inference and use the AutoGuide functionality to define the form of the surrogate posterior, and constrain the variational parameters $\boldsymbol{\zeta}$.

3.1 The Laplace Approximation

The Laplace Approximation approximates a posterior probability distribution as Gaussian centered around the maximum a posteriori (MAP) estimate. Here we approximate the posterior distribution over the parameter vector $\boldsymbol{\phi} = \{\log A_V, \mu + \delta M, \theta_1, \Delta t_{\max}, \boldsymbol{e}\}$ using the Laplace approximation, where a log transform is used to enforce a positivity constraint on A_V. For some multivariate posterior distribution $p(\boldsymbol{\phi}|x)$, with parameters $\boldsymbol{\phi}$ and data x, we can perform a second-order Taylor expansion about the MAP estimate $\hat{\boldsymbol{\phi}}$:

$$\log p(\boldsymbol{\phi}|x) \approx \log p(\hat{\boldsymbol{\phi}}|x) - \frac{1}{2}(\boldsymbol{\phi} - \hat{\boldsymbol{\phi}})^\top \boldsymbol{\Xi} (\boldsymbol{\phi} - \hat{\boldsymbol{\phi}}). \tag{11}$$

Here $\nabla \log p(\boldsymbol{\phi}|x)\big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}} = 0$ by the definition of MAP, and $\boldsymbol{\Xi} = -\nabla\nabla \log p(\boldsymbol{\phi}|x)\big|_{\boldsymbol{\phi}=\hat{\boldsymbol{\phi}}}$ is the Hessian of the negative log-posterior at $\hat{\boldsymbol{\phi}}$ (see e.g. Rasmussen & Williams 2006; Gelman et al. 2014). The posterior distribution can thus be approximated by a Gaussian distribution,

$$q_\zeta(\boldsymbol{\phi}) \approx \mathrm{N}(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}}, \boldsymbol{\Xi}^{-1}). \tag{12}$$

The Laplace Approximation was implemented using the AutoLaplaceApproximation class within NUMPYRO with parameters initialized to the median of their respective priors. The AutoLaplaceApproximation interface is the same as all other AutoGuides in NUMPYRO, but the ELBO is maximized between the true posterior and a Delta function guide to find the MAP estimate.

3.2 The Multivariate Normal Distribution

Here we define the parameter vector $\boldsymbol{\phi} = \{\log A_V, \mu + \delta M, \theta_1, \Delta t_{\max}, \boldsymbol{e}\}$ (again using a log transform to constrain $A_V > 0$) and approximate the joint posterior with a surrogate of the form:

$$q_\zeta(\boldsymbol{\phi}) = \mathrm{N}(\boldsymbol{\phi}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{13}$$

where variational parameters $\boldsymbol{\zeta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The multivariate normal guide with a full covariance matrix was implemented using the AutoMultivariateNormal class within NUMPYRO.

3.3 The Multivariate Zero-Lower-Truncated Normal Distribution

Here we define the generic multivariate Zero-Lower-Truncated Normal (MVZLTN) distribution as a multivariate distribution with the first variable being characterized by a ZLTN distribution (as defined

below) and the remainder following a multivariate normal distribution conditional on the first variable. As described in Section 2.2, we use this distribution to impose a positivity constraint on A_V while maintaining a full covariance matrix between all model parameters $\boldsymbol{\phi} = \{A_V, \mu + \delta M, \theta_1, \Delta t_{\max}, \boldsymbol{e}\}$. To allow for nonzero density at $A_V = 0$, we are not imposing a log transform on A_V.

3.3.1 1D ZLTN Distribution

If a generic one-dimensional truncated normal distribution is truncated on the left at $\phi = a$ and on the right at $\phi = b$, then the Zero-Lower-Truncated Normal (ZLTN) distribution has $a = 0$ and $b = \infty$. The probability density function (PDF) has the form

$$\text{ZLTN}(\phi \mid \mu, \sigma^2) = \begin{cases} \dfrac{1}{\sigma} \dfrac{\psi([\phi - \mu]/\sigma)}{\Psi(\mu/\sigma)} & \text{if } \phi \geq 0 \\ 0 & \text{otherwise,} \end{cases} \quad (14)$$

where μ and σ are the mean and standard deviation of the untruncated distribution, and $\psi(\phi)$ and $\Psi(\phi)$ are respectively the PDF and cumulative density function (CDF) of a standard unit normal distribution.

3.3.2 Multivariate ZLTN Distribution

We define an N-dimensional distribution over $\boldsymbol{\phi} = (\phi_t, \phi_u^1, \ldots, \phi_u^{N-1})^\top$, characterized by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The first variable (ϕ_t) has positive support and is ZLTN distributed. We call this the truncated variable. The remaining $N - 1$ variables, $\boldsymbol{\phi}_u = (\phi_u^1, \ldots, \phi_u^{N-1})^\top$, are untruncated and follow a multivariate normal distribution conditional on ϕ_t. The mean and covariance parameters can be partitioned as:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_t \\ \boldsymbol{\mu}_u \end{bmatrix}, \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_t^2 & \boldsymbol{\Sigma}_{ut}^\top \\ \boldsymbol{\Sigma}_{ut} & \boldsymbol{\Sigma}_{uu} \end{bmatrix} \quad (15)$$

where μ_t is a scalar, $\boldsymbol{\mu}_u$ is a vector of length $N - 1$, σ_t is a scalar, $\boldsymbol{\Sigma}_{ut}$ is a vector of length $N - 1$, and $\boldsymbol{\Sigma}_{uu}$ is an $(N - 1) \times (N - 1)$ matrix. Here μ_t and σ_t^2 are parameters of a ZLTN distribution, and the rest are parameters of a multivariate normal distribution.

First, we define the marginal probability density of the truncated parameter ϕ_t as $\text{ZLTN}(\phi_t \mid \mu_t, \sigma_t^2)$. The conditional distribution of the remaining parameters is then:

$$\boldsymbol{\phi}_u \mid \phi_t \sim N(\boldsymbol{\mu}_{u|t}, \boldsymbol{\Sigma}_{uu|t}) \quad (16)$$

where the conditional mean is

$$\boldsymbol{\mu}_{u|t} = \boldsymbol{\mu}_u + \boldsymbol{\Sigma}_{ut} \frac{\phi_t - \mu_t}{\sigma_t^2} \quad (17)$$

and the conditional covariance is

$$\boldsymbol{\Sigma}_{uu|t} = \boldsymbol{\Sigma}_{uu} - \frac{1}{\sigma_t^2} \boldsymbol{\Sigma}_{ut} \boldsymbol{\Sigma}_{ut}^\top \quad (18)$$

The log probability density of a full N-dimensional parameter vector $\boldsymbol{\phi} = (\phi_t, \boldsymbol{\phi}_u)^\top$ can be calculated as the sum of the log probability density of the truncated parameter, and that of all other variables conditioned on the truncated one:

$$\log \text{MVZLTN}(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$\equiv \log N(\boldsymbol{\phi}_u \mid \boldsymbol{\mu}_{u|t}, \boldsymbol{\Sigma}_{uu|t}) + \log \text{ZLTN}(\phi_t \mid \mu_t, \sigma_t^2) \quad (19)$$

The first term on the right hand side is calculating the log probability density of the multivariate Gaussian distribution in Equation 16 and the second term is the log probability density of a 1D ZLTN distribution (Equation 14). Useful marginal probability densities over subvectors of $\boldsymbol{\phi}$ are derived in Appendix A.

3.3.3 Using Multivariate ZLTN as a Surrogate

In our application, we use the MVZLTN distribution constructed above, with a joint density computed as in Eq. 19, as the surrogate posterior distribution:

$$q_\zeta(\boldsymbol{\phi}) = \text{MVZLTN}(\boldsymbol{\phi} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (20)$$

where $\phi_t = A_V$ is the parameter whose distribution is truncated to be positive. The variational parameters are $\zeta = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ as defined in Equation 15.

3.3.4 Custom AutoGuide Implementation

We first extend the `Distribution` class in NumPyro to create a custom joint MVZLTN distribution with the first variable following a ZLTN distribution and all other variables a multivariate normal distribution conditional on the first . We override the `sample()` and `log_prob()` methods and implement new ones using the equations above. We then extend the `AutoContinous` guide class and create a new type of `AutoGuide` class that has a MVZLTN posterior. Utilizing the `AutoGuide` functionality in NumPyro enables us to fit a MVZLTN posterior to any model.

4 DATA

To test our variational inference implementation of BayeSN, we fit both real and simulated SN Ia light curves.

4.1 Foundation Supernova Survey

As our test dataset, we use the first data release of the Foundation Supernova Survey (Foley et al. 2018; Jones et al. 2019). These data include $griz$ ($\sim$ 3500–9500 Å) light curves of 180 spectroscopically normal SNe Ia observed on the Pan-STARRS telescope. The light curves have a fairly homogeneous time sampling, with observations following a nominal sequence of $(-5, 0, 5, 10, 18, 26, 35)$ days relative to peak brightness (Foley et al. 2018). The filter set, cadence, and telescope used make these data a good representation of other current datasets to which a scalable Variational-BayeSN model will be directly applicable (e.g. the Young Supernova Experiment; Aleo et al. 2023; and Pan-STARRS Medium Deep Survey; Rest et al. 2014; Scolnic et al. 2018; Villar et al. 2020). They are also a valuable precursor to the data expected from LSST. Additionally, previous BayeSN analyses (Thorp et al. 2021; Ward et al. 2023b) have made use of the Foundation data, and they form the training set of the T21 model used throughout this work. We limit ourselves here to the 157 Foundation SNe Ia used in Thorp et al. (2021).

4.2 Simulated Data

As well as analyzing real data from Foundation, we also perform tests using data simulated from the T21 BayeSN forward model. These data are simulated to closely mimic the properties of the real Foundation data, but also give us access to ground truth parameter values. To simulate our data we sampled parameter values from the

following priors[8]:

$$A_v \sim \text{Exponential}(\tau_A = 0.194) \tag{21}$$

$$\theta_1 \sim N(0, 1) \tag{22}$$

$$z \sim U(0.015, 0.08) \tag{23}$$

$$\Delta t_{\max} \sim N(0, 5^2) \tag{24}$$

where the distance modulus μ was calculated from the sampled redshift value z (assuming a flat ΛCDM cosmology with $H_0 = 73.24$ km s^{-1} Mpc^{-1}, $\Omega_0 = 0.28$, $\Omega_\Lambda = 0.72$; Riess et al. 2016). We then used the BAYESN forward model to simulate SEDs and "observed" light curves using these sampled parameters, with simulated photometric errors of 0.05 mag, and a Gaussian error of 0.001 applied to the redshifts to mimic spectroscopic redshift error. In our fits, we fix redshift to this "observed" value, rather than the true z (reflecting the way in which spectroscopic redshifts are treated in real analyses). In practice, this error only affects the redshift and time dilation of the SED model, and the effect is negligible at low-z after integration through the bandpasses and the addition of photometric error.

The simulations followed an idealized version of the Foundation survey strategy, with a cadence of 4 rest-frame days, with the earliest observation at 8 rest-frame days before peak, and the latest at 36 rest-frame days after peak in $griz$ (with observations in the four bands being coincident in time). The nominal observing sequence in the Foundation survey (Foley et al. 2018) was $[-5, 0, 5, 10, 18, 26, 35]$ days relative to peak (in the observer frame), meaning our simulations cover a comparable range of phases and have similar sampling close to the peak of the light curve. Our assumed photometric uncertainties of 0.05 mag are comparable to the mean photometric uncertainty of 0.045 mag for the full Foley et al. (2018) dataset.

5 RESULTS AND VALIDATION

Both variational approximations (MVZLTN and multivariate normal) and the Laplace Approximation are iteratively optimized, where each iteration consists of a single loss calculation and optimizer update step. In all cases, we fit the Laplace Approximation first for 15,000 iterations, which in practice was found to be sufficient for finding the MAP estimate. For the MVZLTN and multivariate normal guides, we then initialize on the median values of $\Delta t_{\max}, \theta_1, A_V$, and μ from the Laplace approximation and train further for 10,000 iterations with the new guide. We use the ADAM optimizer (Kingma & Ba 2015) with a learning rate of 0.01 for the Laplace approximation and 0.005 for the MVZLTN and multivariate normal guides to optimize the ELBO computed with 5 "particles", indicating the number of samples from the surrogate posterior used to approximate the expectations in the ELBO (Eq. 3). After optimizing the variational parameters of the guide for a given supernova, we then draw 1,000 parameter samples from the resulting approximate posterior distribution.

As a point of comparison, we also fit the same data with MCMC using the No-U-Turn Sampling (NUTS) method in NUMPYRO (Hoffman & Gelman 2014). We use four chains with 250 warm-up samples and 250 additional samples in each, as is done by previous BAYESN analyses (Thorp et al. 2021; Thorp & Mandel 2022; Grayling et al. 2024) to ensure a sufficient effective sample size (computed by NUMPYRO based on auto-correlation length of the chains; see Geyer 1992, 2011;

Hoffman & Gelman 2014; Gelman et al. 2014). Post-warmup, the NUTS sampler obtains effective samples at a rate $\approx$ 0.5–2 per iteration. Only the 250 post-warmup samples in each chain are used, yielding 1,000 total posterior samples (with effective sample sizes $\approx$ 400–2,000)[9] for each SN.

5.1 Results on Simulated Data

Figures 1 and 2 show the results of fitting 1000 simulated supernovae with MVZLTN VI and multivariate normal VI compared to the MCMC fits and the true parameter values used to generate the simulated light curves. The top row compares both VI and MCMC fits against the truth, the middle row shows the residuals of both VI and MCMC compared to the truth, and the bottom row shows the residuals between VI and MCMC.

Both the VI and MCMC posterior distributions are able to recover the true parameter values well. The median VI-true residuals (seen in the second row of Figures 1 and 2) standardized by the corresponding VI standard deviations for μ, θ_1, and A_V, respectively are -0.04σ, 0.04σ, and 0.11σ for the MVZLTN posterior and -0.07σ, -0.05σ, and 0.15σ for the multivariate normal posterior. For the MCMC approximation, the corresponding median standardized residuals are -0.04σ, 0.03σ, and 0.03σ for μ, θ_1, and A_V, respectively.

The residuals between the VI and MCMC parameter samples are small compared to the typical errors from the MCMC. The bottom row of Figures 1 and 2 shows the distribution of residual values (VI - MCMC) in the context of representative standard deviations of the MCMC marginal distributions for each of μ, θ_1, and A_V. The median VI - MCMC residuals scaled by the MCMC standard deviations for μ, θ_1, and A_V are 0.002σ, 0.02σ, and 0.06σ for the MVZLTN posterior and -0.01σ, -0.07σ, and 0.07σ for the multivariate normal posterior.

At $A_V \lesssim 0.01$, the VI residuals are almost exclusively positive relative to the truth. This is to be expected as we are constraining the marginal distribution of A_V to be positive, hence for low values of true A_V it is unlikely that the VI fit could underestimate them without being negative. At larger value of true A_V, there is a more even distribution of positive and negative residual values.

For VARIATIONAL-BAYESN, the best-case scenario is that the VI approximation matches the performance of the MCMC, which would indicate no tradeoff between computational efficiency and the quality of the posterior approximation. However, this would also require the chosen form of the surrogate posterior to perfectly match that of the true posterior, which in reality is unlikely to be completely true. Thus, some variance is to be expected both between the MCMC samples and true parameters, and the VI approximation and MCMC samples.

As shown in the last row of Figures 1 and 2 and the numerical summaries presented above, both the MVZLTN and multivariate normal VI approximations slightly overestimate the marginal values of A_V compared to MCMC, while the multivariate normal VI also appears to underestimate θ_1. Meanwhile, in both cases, there are neglible differences between the VI and MCMC estimates of the distance moduli μ. The marginal distributions of the VI approximations will be further validated in Section 5.3.2.

---

[8] We use the notation Exponential(τ) to denote an exponential distribution with a scale parameter τ. Here τ_A is a population mean A_V.

[9] Effective sample sizes larger than the number of posterior samples (i.e. super-efficiency) are possible for NUTS, which can generate chains with negative auto-correlation (see discussion in Stan Development Team 2024; Vehtari et al. 2021).
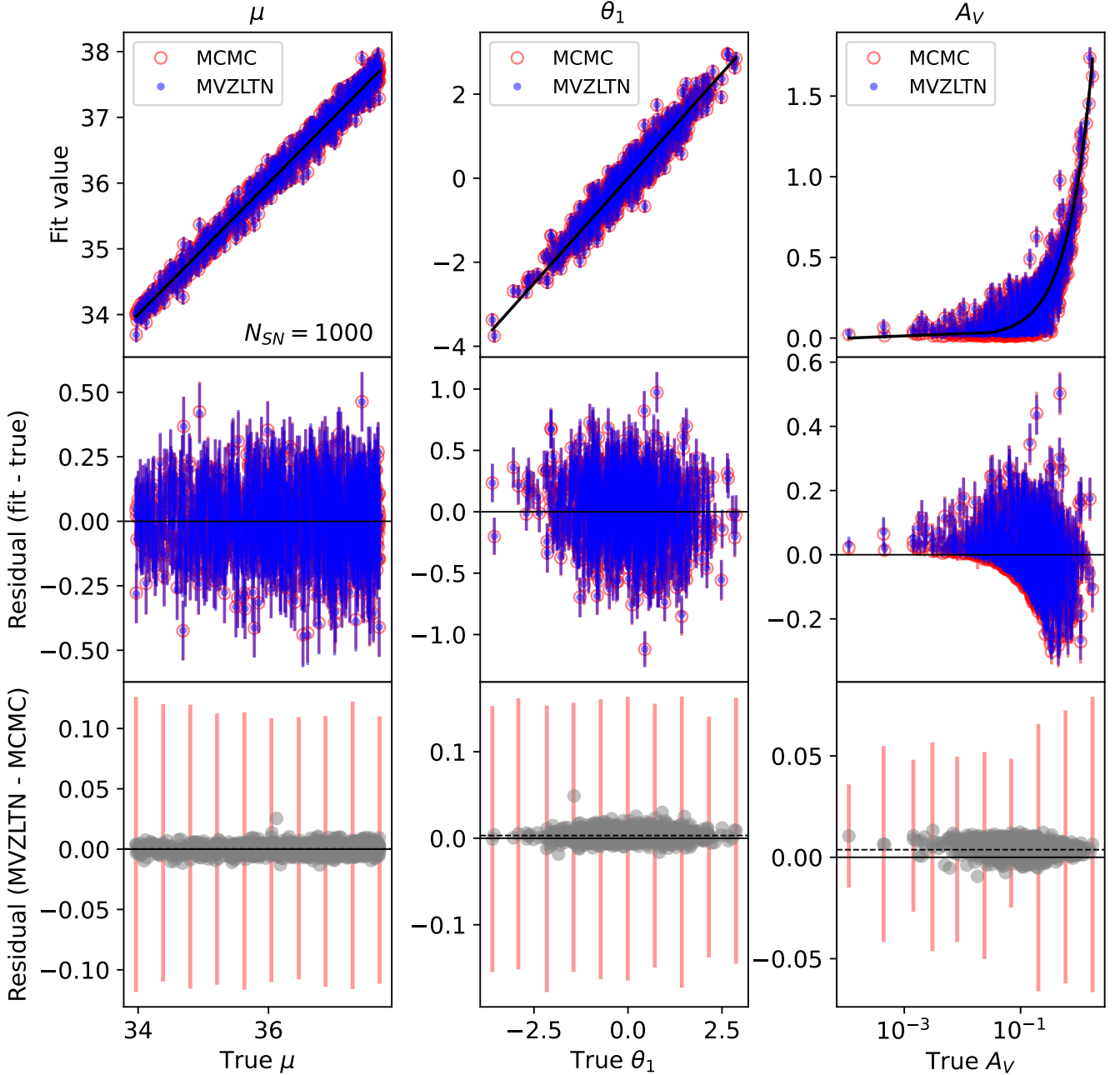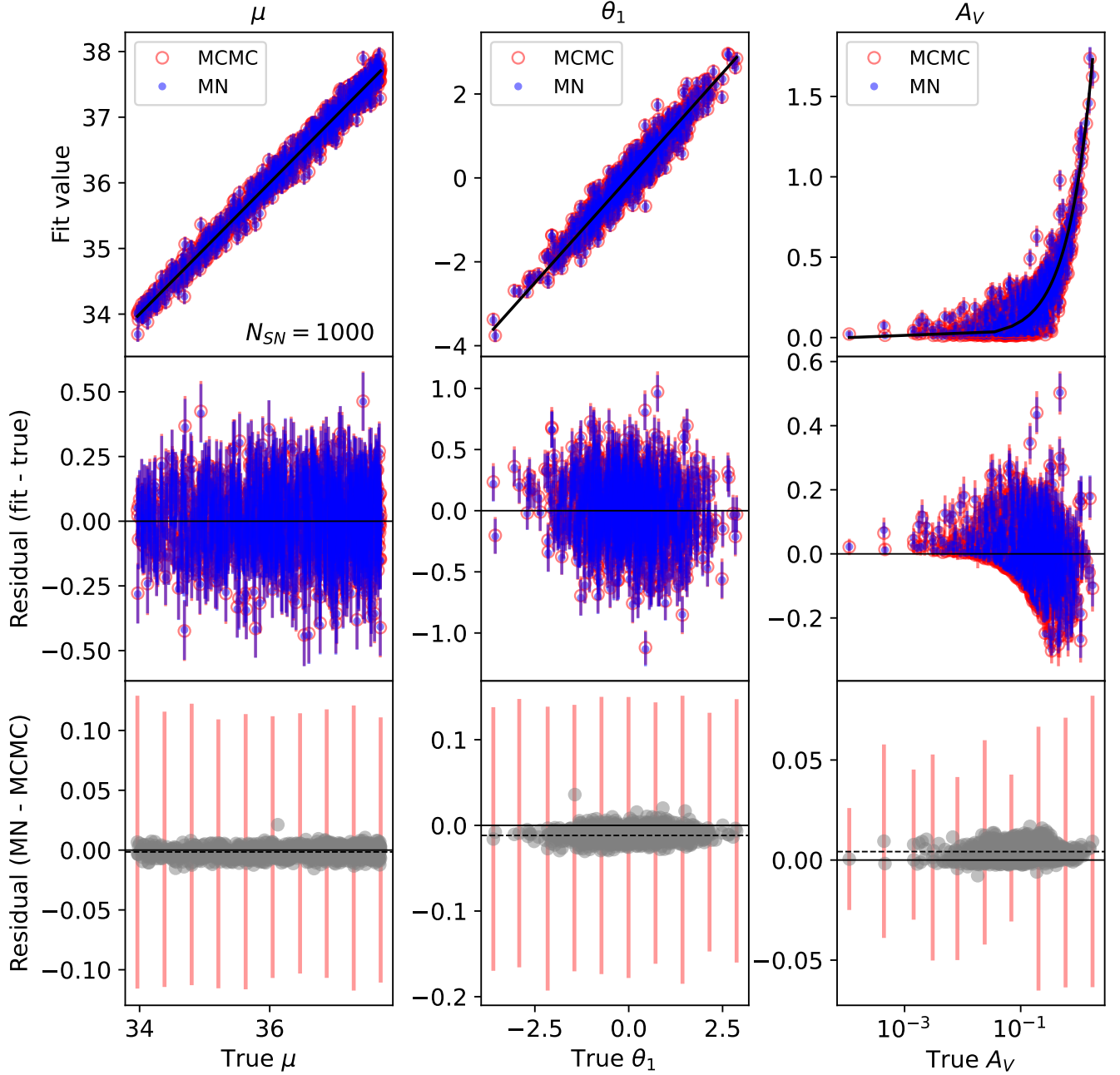
Figure 1. A comparison of MVZLTN VI posterior fits (blue) and MCMC posterior fits (red) with the true parameter values for 1000 simulate SN Ia for μ, θ_1, and A_V. The top row compares the median posterior value vs. the true parameter value, the middle row shows the residuals (fit - true) for MCMC and MVZLTN VI, and the bottom row the residuals (VI - MCMC) as a function of the true parameter value. Black dotted lines in the bottom row denote the median residual (MVZLTN - MCMC) value. The red error bars in the bottom row indicate the typical posterior standard deviation derived from MCMC fits to SNe at a set of 10 representative values spanning each parameter range. Note that the true A_V is plotted on a $\log_{10}$ scale to better show the behavior at low A_V.

5.2 Results on Real Data

We fit the Foundation Dataset (described in Section 4.1), which consists of 157 real SNe Ia. Figure 3 shows the marginal posterior for A_V, θ_1, and μ for SN2017cjv, a low extinction SN Ia, with MCMC, MVZLTN VI, multivariate normal VI, and the Laplace approximation, while Figure 4 shows the same for ASASSN-16ay, a highly extinguished SN Ia. For A_V far from zero as in Figure 4, the

four different types of posterior approximations agree quite well and are virtually indistinguishable.

The top left subplot in Figure 3 shows the marginal distribution of A_V for each of the four posterior approximations. As expected the multivariate normal VI and the Laplace approximation show log-normal behavior with zero density as A_V approaches 0. The MCMC distribution shows non-negligible density as A_V approaches 0, which is most closely matched by the MVZLTN VI posterior. Of

Figure 2. A comparison of multivariate normal (MN) VI posterior fits (blue) and MCMC posterior fits (red) with the true parameter values for 1000 simulated SN Ia for μ, θ_1, and A_V. The top row compares the median posterior value vs. the true parameter value, the middle row shows the residuals (fit - true) for MCMC and multivariate normal VI, and the bottom row the residuals (VI - MCMC) as a function of the true parameter value. Black dotted lines in the bottom row denote the median residual (MN - MCMC) value. The red error bars in the bottom row indicate the typical posterior standard deviation derived from MCMC fits to SNe at a set of 10 representative values spanning each parameter range. Note that the true A_V is plotted on a $\log_{10}$ scale to better show the behavior at low A_V.

the variational approximations, the MVZLTN appears to model a low-extinction SN's posterior the best.

Figure 5 displays a Hubble diagram of all SNe in the Foundation dataset fit with MVZLTN VI. The root-mean squared error (RMSE) is 0.130 mag, which is consistent with that from the same analysis done by Thorp et al. (2021) using HMC. We perform the same analysis with the multivariate normal VI, Laplace Approximation and

MCMC, which yield RMSE values of 0.129, 0.127, and 0.130 mag, respectively.

5.3 Validation of VI Performance

It is possible to optimize the parameters for a surrogate posterior of a given form using VI but still not approximate the true posterior well.

Figure 4. Corner plot showing the posterior distributions from MCMC (red), MVZLTN VI (black), multivariate normal VI (blue), and Laplace Approximation (green), for ASASSN-16ay, a SN with high A_V.
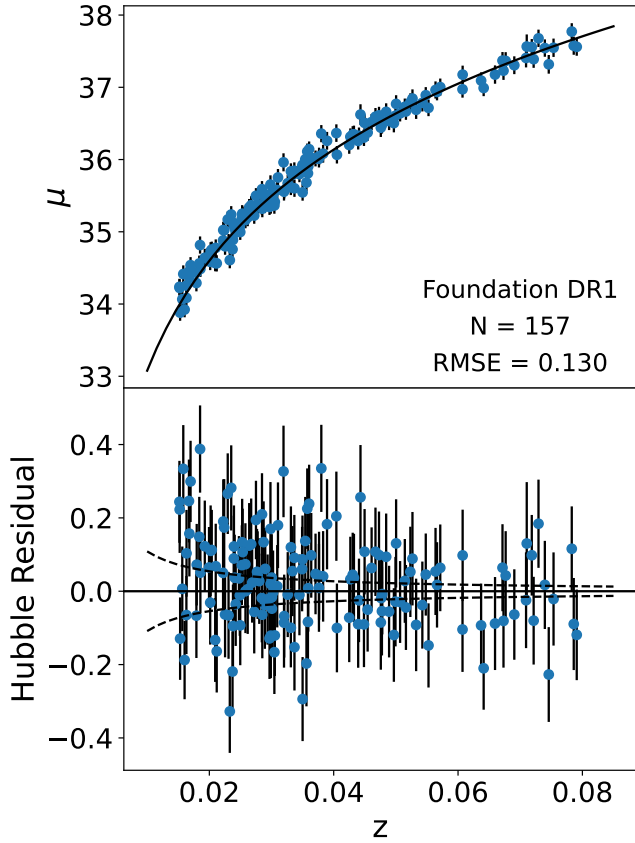
Figure 3. (top) Zoomed-in plot of the estimated marginal posterior distributions over A_V. For the VI and Laplace approximation, we plot the analytic form of the approximate posterior (see Appendix A for the MVZLTN marginals). (bottom) Corner plot showing the posterior distributions from MCMC (red), MVZLTN VI (black), multivariate normal VI (blue), and Laplace Approximation (green), for SN2017cjv, an SN with low A_V.

We use the methods for validating the performance of a VI model presented by Yao et al. (2018) (hereafter Y18) to evaluate the quality of our variational approximations.

5.3.1 *Pareto-Smoothed Importance Sampling (PSIS)*

The Pareto-Smoothed Importance Sampling (PSIS; Vehtari et al. 2024) diagnostic evaluates the goodness of fit of the surrogate posterior distribution compared to the true joint probability (Y18). For each of N posterior samples $\phi_1, \ldots, \phi_N$, the importance ratio r_n is calculated as:

$$r_n = \frac{p(\phi_n, x)}{q_\zeta(\phi_n)} \qquad (25)$$

for data x, parameters ϕ_n, joint density $p(\phi_n, x)$, and surrogate posterior $q_\zeta(\phi_n)$. The M largest importance ratios – where $M = \min(N/5, 3\sqrt{N})$ – are fit to a generalized Pareto distribution yielding a shape parameter $\hat{k}$ (Vehtari et al. 2024). If $\hat{k} < 0.5$, this indicates that the variational approximation is a good approximation to the true joint density, while if $0.5 < \hat{k} < 0.7$, the approximation is still acceptable (Y18). A $\hat{k} > 0.7$ suggests that the variational approximation has large uncertainty (Y18).

In calculating the PSIS diagnostic for our MVZLTN VI fits of the Foundation dataset, we fit the Foundation dataset as described at the beginning of Section 5 and use $N = 5000$ parameter samples from the posterior (using > 2000 samples as suggested by Vehtari et al. 2024) to calculate $\hat{k}$. Figure 6 shows the distribution of $\hat{k}$ values as a function of the difference in the median distance modulus μ from the MVZLTN VI and MCMC fits. The values of the residuals do not seem to correlate with the $\hat{k}$ diagnostic and, save for one outlier, appear approximately normally distributed around a μ residual of zero. Vehtari et al. (2024) note that performance of the PSIS diagnostic deteriorates as the number of parameters in the posterior increases, and Huggins et al. (2020) find that the diagnostic does not always correlate with the quality of the posterior approximation as intended. Overall, we do not find the PSIS $\hat{k}$ diagnostic to be effective in assessing the quality of our variational approximations, at least in terms of producing a distance modulus point estimate; however, we acknowledge the importance of validating the performance of our VI implementation.

5.3.2 *Variational Simulation-Based Calibration (VSBC)*

The Variational Simulation-Based Calibration (VSBC) diagnostics determine whether a point estimate of a single parameter from its marginal distribution from a variational approximation is likely to be biased (Y18).

For our simulated results shown in Section 5.1, we generate 1,000

Figure 6. Median distance modulus μ residual (MVZLTN VI − MCMC) vs. PSIS $\hat{k}$ diagnostic calculated from $N = 5000$ parameter samples for the Foundation dataset. Dotted line indicates $\hat{k} = 0.7$, a threshold for a good variational approximation of the posterior (Vehtari et al. 2024).

Figure 5. (top) Hubble diagram of distance modulus μ vs. redshift z for 157 supernovae in the Foundation dataset fit with MVZLTN VI. Black line indicates expected Hubble relation under flat ΛCDM with $H_0 = 73.24$ km s^{-1} Mpc^{-1} and $\Omega_0 = 0.28$ (Riess et al. 2016). Error bars indicate 1σ for each supernova. The RMSE for this dataset is 0.13. (bottom) Hubble residuals (difference between fitted μ values and expected Hubble relation shown in top panel) 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.

true parameter values from the priors shown in Section 4, and simulate a set of SN Ia *griz* light curves using each set of these parameters, corresponding to a single simulated SN Ia. We then fit the posterior using each of the outlined methods, and generate 1,000 samples from the posterior. To compute the VSBC diagnostics, we calculate $p_{s\phi}$, the probability of a posterior sample of model parameter ϕ from the variational approximation for simulated SN Ia s being greater than the true value of ϕ (drawn from the priors listed in Section 4) that was used in creating the simulated data for SN Ia s. In practice, this is simply the fraction of our 1,000 posterior samples that are below the true simulated parameter value:

$$p_{s\phi} = \frac{\#\,(\text{posterior samples} > \text{true value})}{\#\,\text{posterior samples}} \qquad (26)$$

for simulated SN Ia $s = 1, \ldots, 1000$ and model parameters $\phi \in \{\mu, \theta_1, A_V\}$. We then investigate whether the distributions of these $p_{s\phi}$ values are symmetric across all simulated SNe for each model parameter ϕ. Yao et al. (2018) note that if the distribution is not symmetric, then the point estimate derived from the VI approximation is unlikely to be close to the median value of the true posterior.
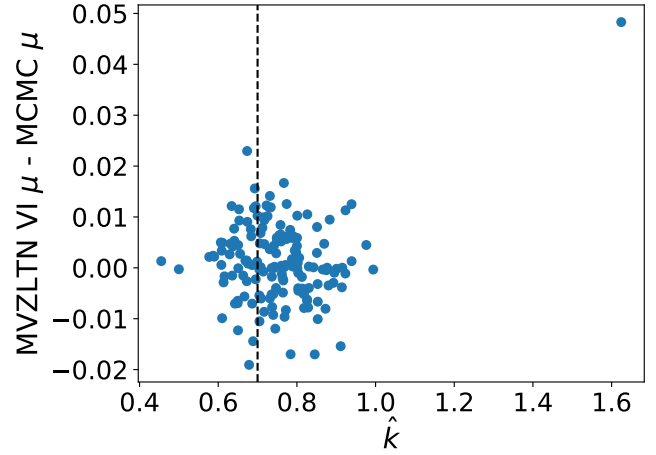
In each panel of Figure 7, we show the histogram of $p_{s\phi}$ for $\phi = \mu$, A_V, or θ_1 across all supernovae $s = 1 \ldots 1000$ for each type of posterior fit. To investigate the symmetry of each marginal distribution, we perform a 2-sided Kolmogorov–Smirnov (KS) test (Kolmogorov 1933; Smirnov 1948; Massey Jr. 1951) on the distributions of p_ϕ and $1 - p_\phi$ as in Y18. The KS test yields a frequentist p-value, here denotes as p_{KS}, which if less than 0.05 suggests that p_ϕ and $1 - p_\phi$ are unlikely to be drawn from the same distribution and thus the distribution of p_ϕ is not symmetric. These p_{KS}-values are annotated on each subplot in Figure 7. Note that this technically violates the requirement of the KS test that the two distributions being compared are independent; however, we chose to follow the procedure for the VSBC diagnostic as described by Y18.

A given marginal distribution from a variational approximation that closely resembles the true posterior would be symmetric about $p_\phi = 0.5$, denoted by a black vertical line in Figure 7. The results of the VSBC diagnostics show that there is a systematic offset in the point estimates from the Laplace approximation for all three parameters. While the two VI methods tend to overestimate A_V, the marginal distributions of θ_1 and μ remain centered around the true values. The VSBC diagnostics find that the MCMC marginal distributions are unbiased, suggesting both VI approximations slightly overestimate A_V when compared to the MCMC, but these offsets are small compared to the MCMC uncertainties, as seen in the bottom row of Figures 1 and 2 and the numerical summaries discussed in Section 5.1.

Whereas the PSIS diagnostic evaluates performance on the 28-dimensional joint posterior, the VSBC diagnostics solely operate on the marginal distribution of one parameter, ignoring other parameters even though they may covary. In practice, the most common application of SNe Ia SED fitting models such as BAYESN is to fit the distance-redshift relation for cosmological parameter estimation. For this, the posterior point estimate of the photometric distance modulus μ marginalized over all other light curve parameters is needed, and Figure 7 shows that these point estimates are unbiased in any of the variational approximations.
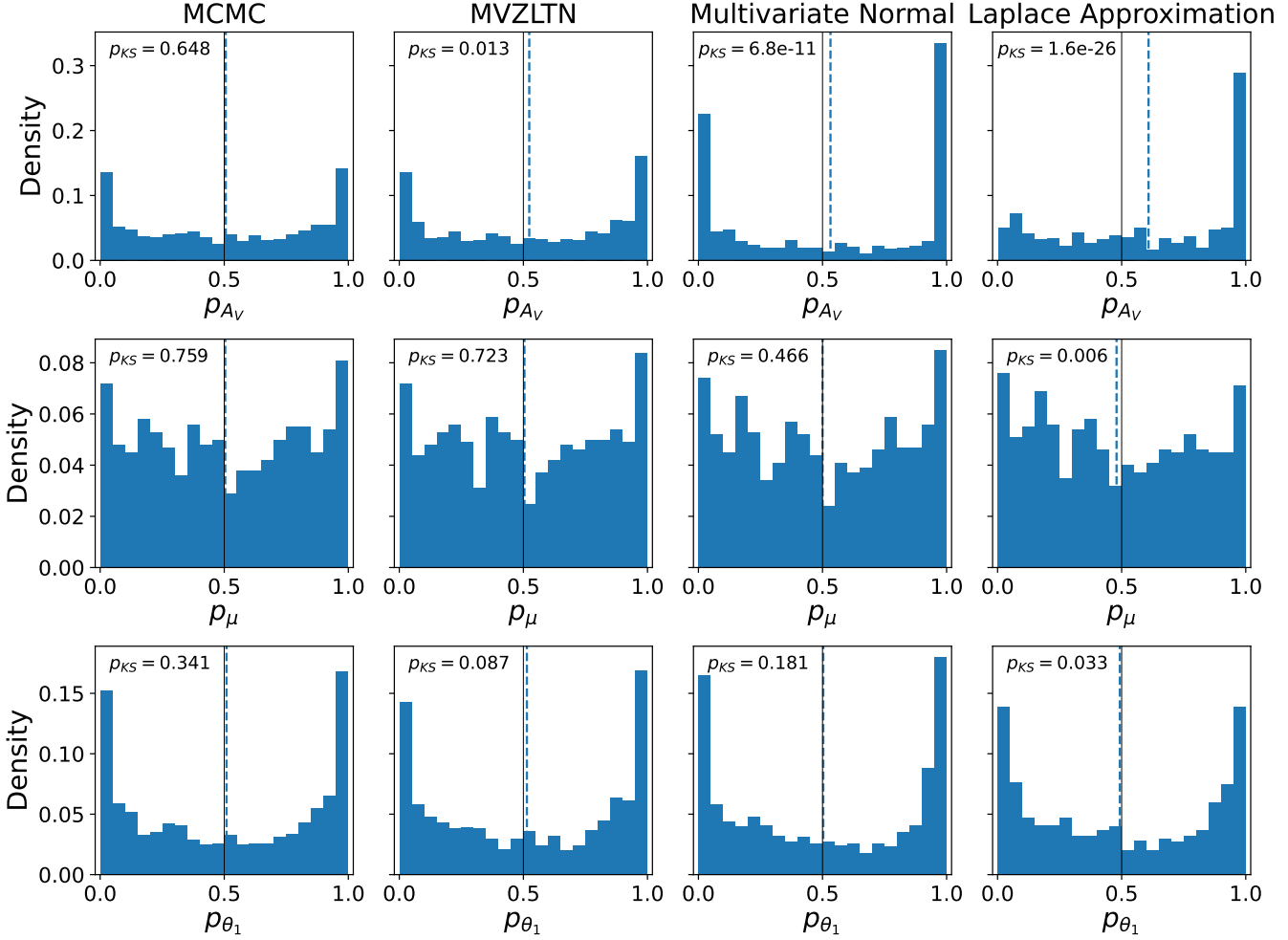
Figure 7. Results from VSBC diagnostic for MCMC, MVZLTN, multivariate normal, and Laplace Approximation posteriors, showing histograms of the probability of a single parameter sample from the posterior being less than the true parameter value, for 1000 simulated SNe. Dotted vertical line indicates mean probability, and solid black line indicates $p = 0.5$ for comparison.

5.4 Computational Performance

The runtime of our NuMPYRO code depends on how many chains and samples are run for MCMC and how many iterations used in VI. We used the `jax.vmap()` method to vectorize the code to fit multiple SNe in parallel (Bradbury et al. 2018). The implementation of BaYeSN in NuMPYRO by Grayling et al. (2024) enables the use of GPUs for further computational speedup.

Table 1 shows the total and per-SN runtimes for fitting the Foundation Dataset, which consists of 157 SNe Ia, and a simulated population of 1000 SNe on both an Apple M2 Pro CPU and an NVIDIA Ampere A100 GPU. On CPU, MCMC was run in parallel across four cores (with one chain per core). In GPU mode, MCMC was run in parallel across four separate GPUs. In contrast, VI required only a single CPU core or GPU. The runtimes in Table 1 show that approximating the posterior with VI is about an order of magnitude faster in addition using less computational resources.

Note that while the per-SN time is listed for comparison between methods, light curves are fit in parallel and the listed time is not representative of the time it would take to fit a single supernova by itself. For the GPUs in particular, the per-SN runtime decreases when fitting larger SN populations. When fitting few SNe, it is likely faster

to use the CPU. Each SN was fit as described at the beginning of Section 5.

6 DISCUSSION

We use both simulated and real data to evaluate the performance of our variational approximations. While the real-world use case of these methods is inferring parameters from observations with no known true values, fitting simulated data where the true parameters are known allows for additional insight into the accuracy and limitations of both VI and MCMC. In practice, a performant VI approximation will yield similar results to MCMC, which our results show on both simulated and real data.

There are several reasons to use VI over MCMC. The first is increased computational efficiency in both time and computing resources. The variational approximation also provides an analytic expression for the posterior distribution, which enables several types of additional analyses, such as calculating importance sampling ratios or the Hessian matrix. An analytic expression for the posterior also enables robust uncertainty propagation, and can be used as a prior in downstream analyses. MCMC, nested sampling, or other

Table 1. A table of runtimes (wall clock time) for fitting the Foundation dataset (157 SNe) and a simulated population (1000 SNe). Per SN runtimes are illustrative[*].

Method	CPU[a]				GPU[b]			
	Foundation		Simulations		Foundation		Simulations	
	total	per SN	total	per SN	total	per SN	total	per SN
MCMC[¶]	3h 48m	1m 27s	16h 24m	1m	4m 42s	2s	11m 20s	0.7s
Laplace Approx.[§]	7m 15s	3s	38m 47s	2s	43s	0.3s	1m 15s	0.1s
MultiNormal VI[§]	26m 37s	10s	2h 24m	9s	42s	0.3s	3m 2s	0.2s
MVZLTN VI[§]	28m 22s	11s	2h 43m	10s	54s	0.3s	2m 45s	0.2s

[a] CPU runtimes are based on an Apple M2 Pro
[b] GPU runtimes are based on an NVIDIA Ampere A100
[¶] MCMC CPU/GPU results used four parallel chains, with one core/GPU per chain
[§] VI / Laplace CPU results were run using a single core
[*] On GPU, the fractional speedup is greater when fitting large batches of SNe. All SNe were fit in parallel, which leads to a considerable performance increase on GPU. However, this is not necessarily the case when running on CPU. For a like-for-like comparison between all methods, we ran parallel fits on both CPU and GPU for this work. For a small sample of objects being fit on CPU, faster per-SN times can be achieved by fitting objects in series, which the public release of BAYESN supports.

posterior sampling methods are likely still the best option for models where the form of the posterior cannot be assumed, is multimodal, or is another form that is difficult to express analytically. However, the compressed representation of the posterior in terms of the variational parameters ζ is also advantageous in terms of data volume, as saving many posterior samples is infeasible in the large-data regime (see e.g. discussion in Malz et al. 2018).

Here we consider a particularly difficult use case with one parameter that is constrained to be nonnegative. The performance of all of the different types of variational posteriors on the non-constrained parameters is quite good, and for most use cases using a Multivariate Gaussian guide or even the Laplace approximation will achieve similar results to MCMC in a fraction of the computational time (as seen in Figure 4) while also providing an analytic posterior and robust uncertainties. This is especially promising in light of the expected quantity of data from the Vera C. Rubin Observatory/LSST in the coming years.

The ZLTN distribution is asymmetric, and so, when using it to approximate the marginal distribution of A_V, the mode and the median are not the same. However, we found in practice that using the median sample as the point estimate led to better results. In calculating point estimates of μ, we marginalize over the other parameters. While it is most important to estimate the distance modulus μ accurately for precision cosmology, this cannot happen without proper modeling of A_V. There is a posterior covariance between μ and A_V as both have a dimming effect. The degeneracy is not perfect, as A_V has a wavelength-dependent effect. Nevertheless, probabilistically modeling the impact of dust on the SN Ia SED is critical to accurate distance estimation.

KL divergence is not symmetric ($KL[P||Q] \neq KL[Q||P]$), and this distinction can have important implications (Margossian et al. 2024). VI uses the ELBO as its loss function, which minimizes the reverse KL divergence $D_{KL}[q_\zeta(\phi)||p(\phi, x)]$ between the true posterior $p(\phi, x)$ and the surrogate posterior $q_\zeta(\phi)$. When approximating a multimodal distribution with a normal distribution, the reverse KL divergence is "mode-locking", meaning that it will perfectly fit one mode of the distribution with zero density around the other modes (Minka 2005; Turner & Sahani 2011). By contrast, the forward KL divergence $D_{KL}[p(\phi, x)||q_\zeta(\phi)]$ is "mode-averaging", meaning that it will approximate a multimodal distribution by spanning its density across all modes and any areas in between modes (Minka 2005;

Turner & Sahani 2011). There are other algorithms, such as expectation propagation, that use the forward KL divergence to avoid getting "stuck" in one mode of the posterior (Minka 2001). In this work, we are not approximating a multimodal posterior distribution, but this is an important consideration for generalized use of VI (Margossian et al. 2024).

7 FUTURE WORK & CONCLUSIONS

In this work we present a VI implementation of BAYESN using NUMPYRO with multivariate normal and MVZLTN surrogate posterior and compare the results on both simulated and real data to MCMC and Laplace Approximation posteriors. We evaluate the PSIS and VSBC diagnostic to validate the quality of our variational approximation, and show that VI can achieve similar results to MCMC but with reduced runtime and computational resources.

All of the code for this work is publicly available at https://github.com/asmuzsoy/bayesn-VI-paper. The code for the custom `AutoGuide` used to fit the MVZLTN surrogate posterior can easily be modified to allow other types of custom posterior forms for a single variable while still maintaining a fully conditional multivariate distribution. The VI implementation has also been incorporated into the larger public BAYESN release by Grayling et al. (2024), available at https://github.com/bayesn/bayesn.

A promising future avenue would be to extend the application of VI to the BAYESN training process, where the model's hyperparameters are inferred. The work we have done here suggests that the possibility for speed-up is large when switching to VI over MCMC. An additional benefit would be that analytic approximate posteriors on the hyperparameters could be easily propagated into downstream analyses. A further extension that would be valuable in the LSST is the inclusion of redshift as a free parameter, so uncertainties associated with photometric host-galaxy redshifts can be marginalized over. This could easily be included in the existing BAYESN framework. Since photometric redshift PDFs (which would be used as a prior in a BAYESN light curve fit) are typically non-analytic and non-Gaussian, they likely be provided in a summarized form such as quantiles (Malz et al. 2018). Frameworks have been developed for

the inclusion of such priors in probabilistic programming languages (e.g. Perepolkin et al. 2023)[10].

There are also alternate methods to potentially better model the dust extinction coefficient A_V. While we model its posterior with a ZLTN distribution with all other variables normal conditional on A_V, the structure of the joint posterior could also be modeled with a copula for a greater deal of flexibility (Tran et al. 2015; Han et al. 2016). Within astronomy, copulas have recently been used by Patil et al. (2023) to model the distribution of stellar chemical abundances in the Milky Way. Modeling the effect of dust extinction is critical to accurate estimation of distances to SN Ia for precision cosmology.

There have been multiple examples of applying Variational Auto-Encoders (VAEs), which use neural networks to learn nonlinear relationships, to extract physical parameters from transient light curves (Boone 2021; Villar et al. 2021). BAYESN explicitly models physical parameters such as the distance modulus μ and dust extinction parameter A_V, and includes additional parameters to model the remaining variance in the SED (see Equation 4). A variation on BAYESN could combine these two approaches, explicitly modeling the SED dependence on physical parameters while allowing a neural network to learn a nonlinear representation of the rest of the model. This would not be as interpretable as the current BAYESN model, but would leverage the nonlinear representation power of neural networks while explicitly maintaining the known key physical parameters.

As the size of astronomical datasets increase and in anticipation of the Rubin Observatory/LSST, variational approximations such as the one in this work could be key to accelerating data analysis pipelines that are currently using slower sampling methods. In the era of precision cosmology, being able to use as much data as possible without draining expensive computational time and resources enables robust estimates and tighter constraints for cosmological parameters.

ACKNOWLEDGEMENTS

[10] See also the earlier approach by B. Goodrich based on Chebyshev polynomials: https://github.com/bgoodri/StanCon2020. Preliminary investigations using the Stan implementation of BAYESN showed this to be viable in an MCMC analysis (S. M. Ward, private communication). This will be integrated into a future release of the NumPyro BAYESN code: https://github.com/bayesn/bayesn/pull/37.

DATA AVAILABILITY

The data for the 180 supernovae in the Foundation DR1 cosmology sample (Foley et al. 2018; Jones et al. 2019) are publicly available at https://github.com/djones1040/Foundation_DR1.

The simulated data shown, as well as all code used in this work, are publicly available on GitHub at https://github.com/asmuzsoy/bayesn-VI-paper.

REFERENCES

Abbott T. M. C., et al., 2024, ApJ, 973, L14
Aleo P. D., et al., 2023, ApJS, 266, 9
Astier P., et al., 2006, A&A, 447, 31
Astropy Collaboration et al., 2013, A&A, 558, A33
Barbary K., 2021, extinction: Dust extinction laws, Astrophysics Source Code Library, record ascl:2102.026
Bingham E., et al., 2019, J. Machine Learning Res., 20, 1
Blei D. M., Kucukelbir A., McAuliffe J. D., 2017, J. American Statistical Association, 112, 859
Boone K., 2021, AJ, 162, 275
Bradbury J., et al., 2018, JAX: composable transformations of Python+NumPy programs, http://github.com/google/jax
Branch D., Tammann G. A., 1992, ARA&A, 30, 359
Brout D., Scolnic D., 2021, ApJ, 909, 26
Brout D., et al., 2019, ApJ, 874, 106
Burns C. R., et al., 2011, AJ, 141, 19
Burns C. R., et al., 2014, ApJ, 789, 32
Commins E. D., 2004, New Astron. Rev., 48, 567
Conley A., et al., 2008, ApJ, 681, 482
Cook S. R., Gelman A., Rubin D. B., 2006, J. Comput. Graphical Statistics, 15, 675
de Soto K. M., et al., 2024, ApJ, 974, 169
Dhawan S., et al., 2022, MNRAS, 510, 2228
Dhawan S., Thorp S., Mandel K. S., Ward S. M., Narayan G., Jha S. W., Chant T., 2023, MNRAS, 524, 235
Draine B. T., 2003, ARA&A, 41, 241
Feroz F., Hobson M. P., Bridges M., 2009, MNRAS, 398, 1601
Fitzpatrick E. L., 1999, PASP, 111, 63
Foley R. J., et al., 2018, MNRAS, 475, 193
Foreman-Mackey D., 2016, J. 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
Friedman A. S., et al., 2015, ApJS, 220, 9
Frieman J. A., et al., 2008, AJ, 135, 338
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
Geyer C. J., 1992, Statistical Sci., 7, 473

Geyer C. J., 2011, in Brooks S., Gelman A., Jones G. L., Meng X.-L., eds, Handbook of Markov Chain Monte Carlo. Chapman & Hall/CRC, pp 3–48, https://www.mcmchandbook.net/HandbookChapter1.pdf

Grayling M., Thorp S., Mandel K. S., Dhawan S., Uzsoy A. S. M., Boyd B. M., Hayes E. E., Ward S. M., 2024, MNRAS, 531, 953

Gunapati G., Jain A., Srijith P. K., Desai S., 2022, Publ. Astron. Soc. Australia, 39, e001

Guy J., Astier P., Nobili S., Regnault N., Pain R., 2005, A&A, 443, 781

Guy J., et al., 2007, A&A, 466, 11

Hamuy M., et al., 1996a, AJ, 112, 2408

Hamuy M., Phillips M. M., Suntzeff N. B., Schommer R. A., Maza J., Smith R. C., Lira P., Aviles R., 1996b, AJ, 112, 2438

Han S., Liao X., Dunson D., Carin L., 2016, in Gretton A., Robert C. C., eds, Proc. Machine Learning Res. Vol. 51, Proceedings of the 19th International Conference on Artificial Intelligence and Statistics. PMLR, Cadiz, Spain, pp 829–838 (arXiv:1506.05860)

Handley W. J., Hobson M. P., Lasenby A. N., 2015, MNRAS, 453, 4384

Harris C. R., et al., 2020, Nature, 585, 357

Hatano K., Branch D., Deaton J., 1998, ApJ, 502, 177

Hicken M., et al., 2012, ApJS, 200, 12

Hinton S., Brout D., 2020, J. Open Source Software, 5, 2122

Hoffman M. D., Gelman A., 2014, J. Machine Learning Res., 15, 1593

Hoffman M. D., Blei D. M., Wang C., Paisley J., 2013, J. Machine Learning Res., 14, 1303

Holwerda B. W., Reynolds A., Smith M., Kraan-Korteweg R. C., 2015, MNRAS, 446, 3768

Hounsell R., et al., 2018, ApJ, 867, 23

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

Huggins J., Kasprzak M., Campbell T., Broderick T., 2020, in Chiappa S., Calandra R., eds, Proc. Machine Learning Res. Vol. 108, Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. PMLR, pp 1792–1802 (arXiv:1910.04102)

Hunter J. D., 2007, Comput. Sci. Eng., 9, 90

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

Jha S., Riess A. G., Kirshner R. P., 2007, ApJ, 659, 122

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

Jones D. O., et al., 2022, ApJ, 933, 172

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

Kantorovich L., Rubinstein G. S., 1958, Vestnik Leningrad Univ., 13, 52

Karchev K., Trotta R., Weniger C., 2023, in Machine Learning and the Physical Sciences Workshop, 37th Conference on Neural Information Processing Systems (NeurIPS). (arXiv:2311.15650)

Karchev K., Grayling M., Boyd B. M., Trotta R., Mandel K. S., Weniger C., 2024, MNRAS, 530, 3881

Kenworthy W. D., et al., 2021, ApJ, 923, 265

Kessler R., Scolnic D., 2017, ApJ, 836, 56

Kessler R., et al., 2009a, PASP, 121, 1028

Kessler R., et al., 2009b, ApJS, 185, 32

Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations (ICLR). San Diego, CA, USA (arXiv:1412.6980)

Kingma D. P., Welling M., 2013, preprint, (arXiv:1312.6114)

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

Kolmogorov A., 1933, G. Istituto Ital. Attuari, 4, 83

Krisciunas K., Hastings N. C., Loomis K., McMillan R., Rest A., Riess A. G., Stubbs C., 2000, ApJ, 539, 658

Krisciunas K., et al., 2007, AJ, 133, 58

Kucukelbir A., Tran D., Ranganath R., Gelman A., Blei D. M., 2017, J. Machine Learning Res., 18, 1

Kullback S., Leibler R. A., 1951, Ann. Math. Statistics, 22, 79

Kumar R., Carroll C., Hartikainen A., Martin O., 2019, J. Open Source Software, 4, 1143

LSST Dark Energy Science Collaboration et al., 2018, preprint, (arXiv:1809.01669)

Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511

Liu R., McAuliffe J. D., Regier J., LSST Dark Energy Science Collaboration 2023, J. Machine Learning Res., 24, 1

Malz A. I., Marshall P. J., DeRose J., Graham M. L., Schmidt S. J., Wechsler R., (LSST Dark Energy Science Collaboration 2018, AJ, 156, 35

Mandel K. S., 2011, PhD thesis, Astronomy Department, Harvard Univ.

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., Scolnic D. M., Shariff H., Foley R. J., Kirshner R. P., 2017, ApJ, 842, 93

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

Margossian C. C., Pillaud-Vivien L., Saul L. K., 2024, preprint, (arXiv:2403.13748)

Massey Jr. F. J., 1951, J. American Statistical Association, 46, 68

Minka T. P., 2001, in Breese J. S., Koller D., eds, UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001. Morgan Kaufmann, pp 362–369 (arXiv:1301.2294)

Minka T., 2005, Technical Report MSR-TR-2005-173, Divergence measures and message passing. Microsoft Research, https://tminka.github.io/papers/message-passing/

Papamakarios G., Pavlakou T., Murray I., 2017, in Guyon I., Luxburg U. V., Bengio S., Wallach H., Fergus R., Vishwanathan S., Garnett R., eds, Advances in Neural Information Processing Systems Vol. 30, NIPS 2017. Curran Associates, Inc. (arXiv:1705.07057)

Patil A. A., Bovy J., Jaimungal S., Frankel N., Leung H. W., 2023, MNRAS, 526, 1997

Perepolkin D., Goodrich B., Sahlin U., 2023, Comput. Statistics Data Analysis, 187, 107795

Perlmutter S., et al., 1999, ApJ, 517, 565

Peterson E. R., et al., 2023, MNRAS, 522, 2478

Phan D., Pradhan N., Jankowiak M., 2019, preprint, (arXiv:1912.11554)

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

Phillips M. M., Lira P., Suntzeff N. B., Schommer R. A., Hamuy M., Maza J., 1999, AJ, 118, 1766

Phillips M. M., et al., 2019, PASP, 131, 014001

Pierel J. D. R., et al., 2024, ApJ, 967, 50

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

Popovic B., Brout D., Kessler R., Scolnic D., Lu L., 2021, ApJ, 913, 49

Pskovskii I. P., 1977, Soviet Ast., 21, 675

Ranganath R., Gerrish S., Blei D., 2014, in Kaski S., Corander J., eds, Proc. Machine Learning Res. Vol. 33, Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics. PMLR, Reykjavik, Iceland, pp 814–822 (arXiv:1401.0118)

Rasmussen C. E., Williams C. K. I., 2006, Gaussian Processes for Machine Learning. Adaptive Computation & Machine Learning, MIT Press

Regier J., Miller A. C., Schlegel D., Adams R. P., McAuliffe J. D., Prabhat 2019, Ann. Applied Statistics, 13, 1884

Rest A., et al., 2014, ApJ, 795, 44

Rezende D., Mohamed S., 2015, in Bach F., Blei D., eds, Proc. Machine Learning Res. Vol. 37, Proceedings of the 32nd International Conference on Machine Learning. PMLR, Lille, France, pp 1530–1538 (arXiv:1505.05770)

Riello M., Patat F., 2005, MNRAS, 362, 671

Riess A. G., Press W. H., Kirshner R. P., 1995, ApJ, 438, L17

Riess A. G., Press W. H., Kirshner R. P., 1996a, ApJ, 473, 88

Riess A. G., Press W. H., Kirshner R. P., 1996b, ApJ, 473, 588

Riess A. G., et al., 1998, AJ, 116, 1009

Riess A. G., et al., 1999, AJ, 117, 707

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

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

Rizzato M., Sellentin E., 2023, MNRAS, 521, 1152

Rose B. M., et al., 2021, preprint, (arXiv:2111.03081)

Rust B. W., 1975, Bull. American Astron. Soc., 7, 236

Salim S., Narayanan D., 2020, ARA&A, 58, 529

Sánchez A., Cabrera G., Huijse P., Förster F., 2021, in Machine Learning and the Physical Sciences Workshop, 35th Conference on Neural Information

Processing Systems (NeurIPS). https://ml4physicalsciences.github.io/2021/files/NeurIPS_ML4PS_2021_10.pdf

Sandage A., Tammann G. A., 1993, ApJ, 415, 1

Saunders C., et al., 2018, ApJ, 869, 167

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

Scolnic D. M., et al., 2018, ApJ, 859, 101

Smirnov N., 1948, Ann. Math. Statistics, 19, 279

Speagle J. S., 2020, MNRAS, 493, 3132

Stan Development Team 2024, Stan Modelling Language Users Guide and Reference Manual v.2.34. https://mc-stan.org

Talts S., Betancourt M., Simpson D., Vehtari A., Gelman A., 2018, preprint, (arXiv:1804.06788)

Thorp S., Mandel K. S., 2022, MNRAS, 517, 2360

Thorp S., Mandel K. S., Jones D. O., Ward S. M., Narayan G., 2021, MNRAS, 508, 4310

Thorp S., Mandel K. S., Jones D. O., Kirshner R. P., Challis P. M., 2024, MNRAS, 530, 4016

Tran D., Blei D., Airoldi E. M., 2015, in Cortes C., Lawrence N., Lee D., Sugiyama M., Garnett R., eds, Advances in Neural Information Processing Systems Vol. 28, NIPS 2015. Curran Associates, Inc. (arXiv:1506.03159)

Tripp R., 1998, A&A, 331, 815

Turner R. E., Sahani M., 2011, in Barber D., Cemgil A. T., Chiappa S., eds, Bayesian Time Series Models. Cambridge University Press, pp 104–124, https://doi.org/10.1017/CBO9780511984679.006

Uzsoy A. S., 2022, Master's thesis, Univ. Cambridge, https://www.mlmi.eng.cam.ac.uk/files/2021-2022_dissertations/scalable_bayesian_inference_for_probabilistic_spectrotemporal_models.pdf

Vaserstein L. N., 1969, Problemy Peredachi Inf., 5, 64

Vehtari A., Gelman A., Simpson D., Carpenter B., Bürkner P.-C., 2021, Bayesian Analysis, 16, 667

Vehtari A., Simpson D., Gelman A., Yao Y., Gabry J., 2024, J. Machine Learning Res., 25, 1

Villar V. A., 2022, in Machine Learning and the Physical Sciences Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS). (arXiv:2211.04480)

Villar V. A., et al., 2020, ApJ, 905, 94

Villar V. A., Cranmer M., Berger E., Contardo G., Ho S., Hosseinzadeh G., Lin J. Y.-Y., 2021, ApJS, 255, 24

Virtanen P., et al., 2020, Nature Methods, 17, 261

Wainwright M. J., Jordan M. I., 2008, Foundations Trends Machine Learning, 1, 1

Ward S. M., Dhawan S., Mandel K. S., Grayling M., Thorp S., 2023a, MNRAS, 526, 5715

Ward S. M., et al., 2023b, ApJ, 956, 111

Wingate D., Weber T., 2013, preprint, (arXiv:1301.1299)

Wojtak R., Hjorth J., Hjortlund J. O., 2023, MNRAS, 525, 5187

Yao Y., Vehtari A., Simpson D., Gelman A., 2018, in Dy J., Krause A., eds, Proc. Machine Learning Res. Vol. 80, Proceedings of the 35th International Conference on Machine Learning. PMLR, pp 5581–5590 (arXiv:1802.02538)

APPENDIX A: MARGINAL DENSITIES OF THE MVZLTN

In this appendix, we present analytic results for the marginal probability densities of subsets of $M < N$ variables of the N-dimensional MVZLTN distribution defined in Section 3.3.2. For the results in this section, we will consider an N-vector $\boldsymbol{\phi} = (\phi_t, \boldsymbol{\phi}_u)^\top$, with truncated variable ϕ_t, and $N - 1$ untruncated variables $\boldsymbol{\phi}_u$. The full $\boldsymbol{\phi}$ vector will have an MVZLTN density,

$$P(\boldsymbol{\phi}) = \text{MVZLTN}(\boldsymbol{\phi}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$
$$= \text{ZLTN}(\phi_t | \mu_t, \sigma_t^2) \times \text{N}(\boldsymbol{\phi}_u | \boldsymbol{\mu}_{u|t}, \boldsymbol{\Sigma}_{uu|t}), \tag{A1}$$

with the mean and covariance parameters partitioned as in Equation 15, and with the conditional mean and covariance of $\boldsymbol{\phi}_u|\phi_t$ given by Equations 17 and 18.

In the following subsections, we will derive all univariate and bivariate marginals. We have validated numerical implementations of these analytic expressions against Monte Carlo samples from the MVZLTN joint distribution. Marginal densities over $2 < M < N$ variables are straightforward generalizations of these results.

A1 Univariate Marginal of the Truncated Variable

By definition, the marginal density of the truncated variable ϕ_t is just a univariate ZLTN distribution,

$$P(\phi_t) = \text{ZLTN}(\phi_t|\mu_t, \sigma_t^2), \tag{A2}$$

as defined in Equation 14.

A2 Bivariate Marginal of the Truncated Variable and an Untruncated Variable

Marginalization over any number of untruncated variables can be achieved trivially, and yields a bivariate ZLTN. Marginalizing over all but the ith untruncated variables gives

$$P(\phi_t, \phi_u^i) = \text{ZLTN}(\phi_t | \mu_t, \sigma_t^2) \times \text{N}(\phi_u^i | \mu_{u|t}^i, \Sigma_{uu|t}^{i,i}), \tag{A3}$$

where in the second term we have picked out the ith element of the conditional mean (Eq. 17) and the (i, i)th element of the conditional covariance matrix (Eq. 18).

A3 Univariate Marginal of an Untruncated Variable

From Eq. A3, we can derive the marginal of the ith untruncated variable, ϕ_u^i, by integrating over the truncated variable, ϕ_t. We begin by expanding the first term in Eq. A3 using the definition of the ZLTN (Eq. 14), expanding $\mu_{u|t}^i$ using Eq. 17, and rewriting the second term as a Gaussian density over ϕ_t:

$$P(\phi_u^i) = \int_0^\infty \text{ZLTN}(\phi_t | \mu_t, \sigma_t^2) \times \text{N}(\phi_u^i | \mu_{u|t}^i, \Sigma_{uu|t}^{i,i}) \, d\phi_t$$
$$= \frac{1}{\Psi(\mu_t/\sigma_t)} \int_0^\infty \text{N}(\phi_t | \mu_t, \sigma_t^2) \times \text{N}(\phi_u^i | \mu_{u|t}^i, \Sigma_{uu|t}^{i,i}) \, d\phi_t \tag{A4}$$

First, we handle the case of $\Sigma_{ut}^i \neq 0$. The preceding expression can be written as:

$$P(\phi_u^i) = \frac{1}{\Psi(\mu_t/\sigma_t)} \int_0^\infty \text{N}(\phi_t | \mu_t, \sigma_t^2) \times \frac{\sigma_t^2}{\Sigma_{ut}^i} \text{N}(y | \phi_t, s_t^2) \, d\phi_t, \tag{A5}$$

where $y = \mu_t + \sigma_t^2 \times (\phi_u^i - \mu_u^i)/\Sigma_{ut}^i$ and $s_t^2 = \Sigma_{uu|t}^{i,i} \times \sigma_t^4/(\Sigma_{ut}^i)^2$. We take the product of the two Gaussian densities, giving

$$P(\phi_u^i) = \frac{\sigma_t^2 \times \text{N}(y | \mu_t, \sigma_t^2 + s_t^2)}{\Sigma_{ut}^i \times \Psi(\mu_t/\sigma_t)} \int_0^\infty \text{N}(\phi_t | \tilde{\mu}_t, \tilde{\sigma}_t^2) \, d\phi_t, \tag{A6}$$

where $\tilde{\mu}_t = \tilde{\sigma}_t^2(\sigma_t^{-2}\mu_t + s_t^{-2}y)$ and $\tilde{\sigma}_t^2 = [\sigma_t^{-2} + s_t^{-2}]^{-1}$. Notice that the variable of interest ϕ_u^i is contained in y and therefore $\tilde{\mu}_t$. The remaining integral will then have the solution $[1 - \Psi(-\tilde{\mu}_t/\tilde{\sigma}_t)] = \Psi(\tilde{\mu}_t/\tilde{\sigma}_t)$, giving

$$P(\phi_u^i) = \frac{\sigma_t^2 \times \Psi(\tilde{\mu}_t/\tilde{\sigma}_t)}{\Sigma_{ut}^i \times \Psi(\mu_t/\sigma_t)} \times \text{N}(y | \mu_t, \sigma_t^2 + s_t^2). \tag{A7}$$

In the case of $\Sigma^i_{ut} = 0$, the marginalization simplifies considerably to:

$$P(\phi^i_u) = N(\phi^i_u \mid \mu^i_u, \Sigma^{i,i}_{uu}). \tag{A8}$$

A4 Bivariate Marginal of Two Untruncated Variables

Finally, we will derive the bivariate joint density over the ith and jth untruncated variables, which we will include in a reduced untruncated vector $\boldsymbol{\phi}'_u = (\phi^i_u, \phi^j_u)^\top$. Similarly, we will let $\boldsymbol{\mu}'_u = (\mu^i_u, \mu^j_u)^\top$ and $\Sigma'_{ut} = (\Sigma^i_{ut}, \Sigma^j_{ut})^\top$, and will define a 2×2 submatrix of the full conditional covariance,

$$\Sigma'_{uu|t} = \begin{pmatrix} \Sigma^{i,i}_{uu|t} & \Sigma^{j,i}_{uu|t} \\ \Sigma^{i,j}_{uu|t} & \Sigma^{j,j}_{uu|t} \end{pmatrix}. \tag{A9}$$

Analogously to Equation A3, the trivariate marginal density over ϕ_t and $\boldsymbol{\phi}'_u$ will be given by

$$P(\phi_t, \boldsymbol{\phi}'_u) = \mathrm{ZLTN}(\phi_t \mid \mu_t, \sigma^2_t) \times N\left(\boldsymbol{\phi}'_u \,\middle|\, \boldsymbol{\mu}'_u + \Sigma'_{ut} \frac{\phi_t - \mu_t}{\sigma^2_t}, \Sigma'_{uu|t}\right). \tag{A10}$$

First, we handle the case of $\Sigma'_{ut} \neq 0$. The preceding expression can be written as:

$$P(\phi_t, \boldsymbol{\phi}'_u) = \mathrm{ZLTN}(\phi_t \mid \mu_t, \sigma^2_t) \times N(\boldsymbol{y} \mid \boldsymbol{x}\phi_t, \boldsymbol{W}), \tag{A11}$$

where we have explicitly expanded out the conditional mean (Eq. 17), and define for convenience: $\boldsymbol{y} = \boldsymbol{\phi}'_u - \boldsymbol{\mu}'_u + \Sigma'_{ut}\mu_t/\sigma^2_t$, $\boldsymbol{x} = \Sigma'_{ut}/\sigma^2_t$, and $\boldsymbol{W} = \Sigma'_{uu|t}$. We can recognise the second term in Equation A11 as being the likelihood for ϕ_t in a linear model, $\boldsymbol{y} = \phi_t \boldsymbol{x}$, with correlated Gaussian errors on $\boldsymbol{y}$ with covariance $\boldsymbol{W}$ (see e.g. Gelman et al. 2014, eq. 14.11). We can separate out the ϕ_t-dependent part of this term to obtain

$$N(\boldsymbol{y} \mid \boldsymbol{x}\phi_t, \boldsymbol{W}) = \sqrt{2\pi s^2_t} \times N(\phi_t \mid \hat{\phi}_t, s^2_t) \times N(\boldsymbol{y} \mid \boldsymbol{x}\hat{\phi}_t, \boldsymbol{W}), \tag{A12}$$

where $\hat{\phi}_t = s^2_t \boldsymbol{x}^\top \boldsymbol{W}^{-1} \boldsymbol{y}$ and $s^2_t = (\boldsymbol{x}^\top \boldsymbol{W}^{-1} \boldsymbol{x})^{-1}$ (see Mandel 2011, appendix B5), provided that these inverses are available. In the linear regression analogy, $\hat{\phi}_t$ and s^2_t are the posterior mean and variance of ϕ_t under a flat improper prior (Gelman et al. 2014, §14.7). The ϕ_t-dependent part of Equation A11 can then be simplified into a product of two Gaussian densities, and integrated over ϕ_t:

$$\begin{aligned} P(\boldsymbol{\phi}'_u) &\propto \int_0^\infty \mathrm{ZLTN}(\phi_t \mid \mu_t, \sigma^2_t) \times N(\phi_t \mid \hat{\phi}_t, s^2_t)\, d\phi_t \\ &= \frac{1}{\Psi(\mu_t/\sigma_t)} \int_0^\infty N(\phi_t \mid \mu_t, \sigma^2_t) \times N(\phi_t \mid \hat{\phi}_t, s^2_t)\, d\phi_t \\ &= \frac{N(\mu_t \mid \hat{\phi}_t, \sigma^2_t + s^2_t)}{\Psi(\mu_t/\sigma_t)} \int_0^\infty N(\phi_t \mid \tilde{\phi}_t, \tilde{\tau}^2_t)\, d\phi_t \\ &= \frac{\Psi(\tilde{\phi}_t/\tilde{\tau}_t)}{\Psi(\mu_t/\sigma_t)} \times N(\mu_t \mid \hat{\phi}_t, \sigma^2_t + s^2_t). \end{aligned} \tag{A13}$$

Here, we have defined for convenience $\tilde{\tau}^2_t = [\sigma^{-2}_t + s^{-2}_t]^{-1}$ and $\tilde{\phi}_t = \tilde{\tau}^2_t(\sigma^{-2}_t \mu_t + s^{-2}_t \hat{\phi}_t)$ when taking the product of the two Gaussian densities. The three steps taken in Equation A13 correspond closely to those taken in Equations A5–A7 when deriving the univariate marginal of ϕ^i_u. We can reintroduce the terms from Eq. A12 that we omitted in Eq. A13 to write the full marginal density of $\boldsymbol{\phi}'_u$ as

$$P(\boldsymbol{\phi}'_u) = \sqrt{2\pi s^2_t} \frac{\Psi(\tilde{\phi}_t/\tilde{\tau}_t)}{\Psi(\mu_t/\sigma_t)} \times N(\mu_t \mid \hat{\phi}_t, \sigma^2_t + s^2_t) \times N(\boldsymbol{y} \mid \boldsymbol{x}\hat{\phi}_t, \boldsymbol{W}), \tag{A14}$$

where it is important to remember that $\boldsymbol{y}$, $\hat{\phi}_t$, and $\tilde{\phi}_t$ are all functions of the two untruncated variables of interest $\boldsymbol{\phi}'_u = (\phi^i_u, \phi^j_u)^\top$.

In the special case of $\Sigma'_{ut} = \boldsymbol{0}$, the marginalization over ϕ_t simplifies considerably:

$$P(\boldsymbol{\phi}'_u) = N(\boldsymbol{\phi}'_u \mid \boldsymbol{\mu}'_u, \Sigma'_{uu}), \tag{A15}$$

where Σ'_{uu} is the $[i, j], [i, j]$ 2×2 submatrix of Σ_{uu}.

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