Journal of the American Statistical Association

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Frequentist Consistency of Variational Bayes

Yixin Wang & David M. Blei

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2019, VOL. 114, NO. 527, 1147–1161: Theory and Methods
https://doi.org/10.1080/01621459.2018.1473776

Taylor & Francis
Taylor & Francis Group

Frequentist Consistency of Variational Bayes

Yixin Wang[a] and David M. Blei[b]

[a]Department of Statistics, Columbia University, New York, NY; [b]Department of Statistics and Department of Computer Science, Columbia University, New York, NY

ABSTRACT
A key challenge for modern Bayesian statistics is how to perform scalable inference of posterior distributions. To address this challenge, variational Bayes (VB) methods have emerged as a popular alternative to the classical Markov chain Monte Carlo (MCMC) methods. VB methods tend to be faster while achieving comparable predictive performance. However, there are few theoretical results around VB. In this article, we establish frequentist consistency and asymptotic normality of VB methods. Specifically, we connect VB methods to point estimates based on variational approximations, called frequentist variational approximations, and we use the connection to prove a variational Bernstein–von Mises theorem. The theorem leverages the theoretical characterizations of frequentist variational approximations to understand asymptotic properties of VB. In summary, we prove that (1) the VB posterior converges to the Kullback–Leibler (KL) minimizer of a normal distribution, centered at the truth and (2) the corresponding variational expectation of the parameter is consistent and asymptotically normal. As applications of the theorem, we derive asymptotic properties of VB posteriors in Bayesian mixture models, Bayesian generalized linear mixed models, and Bayesian stochastic block models. We conduct a simulation study to illustrate these theoretical results. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received May 2017
Revised November 2017

KEYWORDS
Asymptotic normality; Bayesian inference; Bernstein–von Mises; Consistency; Statistical computing; Variational methods

1. Introduction

Bayesian modeling is a powerful approach for discovering hidden patterns in data. We begin by setting up a probability model of latent variables and observations. We incorporate prior knowledge by setting priors on latent variables and a functional form of the likelihood. Finally, we infer the posterior, the conditional distribution of the latent variables given the observations.

For many modern Bayesian models, exact computation of the posterior is intractable and statisticians must resort to approximate posterior inference. For decades, Markov chain Monte Carlo (MCMC) sampling (Hastings 1970; Gelfand and Smith 1990; Robert and Casella 2004) has maintained its status as the dominant approach to this problem. MCMC algorithms are easy to use and theoretically sound. In recent years, however, data sizes have soared. This challenges MCMC methods, for which convergence can be slow, and calls upon scalable alternatives. One popular class of alternatives is variational Bayes (VB) methods.

To describe VB, we introduce notation for the posterior inference problem. Consider observations $x = x_{1:n}$. We posit local latent variables $z = z_{1:n}$, one per observation, and global latent variables $\theta = \theta_{1:d}$. This gives a joint,

$$p(\theta, x, z) = p(\theta) \prod_{i=1}^{n} p(z_i \,|\, \theta) p(x_i \,|\, z_i, \theta). \tag{1}$$

The posterior inference problem is to calculate the posterior $p(\theta, z \,|\, x)$.

This division of latent variables is common in modern Bayesian statistics. (In particular, our results are applicable to general models with *local* and *global* latent variables (Hoffman et al. 2013). The number of local variables z increases with the sample size n; the number of global variables θ does not. We also note that the conditional independence of Equation (1) is not necessary for our results. But we use this common setup to simplify the presentation.) In the Bayesian Gaussian mixture model (GMM) (Roberts et al. 1998), the component means, covariances, and mixture proportions are global latent variables; the mixture assignments of each observation are local latent variables. In the Bayesian generalized linear mixed model (GLMM) (Breslow and Clayton 1993), the intercept and slope are global latent variables; the group-specific random effects are local latent variables. In the Bayesian stochastic block model (SBM) (Hofman and Wiggins 2008), the cluster assignment probabilities and edge probabilities matrix are two sets of global latent variables; the node-specific cluster assignments are local latent variables. In the latent Dirichlet allocation (LDA) model (Blei, Ng, and Jordan 2003), the topic-specific word distributions are global latent variables; the document-specific topic distributions are local latent variables. We will study all these examples below.

VB methods formulate posterior inference as an optimization (Jordan et al. 1999; Wainwright and Jordan 2008; Blei, Kucukelbir, and McAuliffe 2016). We consider a family of distributions of the latent variables and then find the member of that family that is closest to the posterior.

CONTACT Yixin Wang *yixin.wang@columbia.edu* Department of Statistics, Columbia University, New York, NY 10027.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JASA*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

Here, we focus on mean-field variational inference (though our results apply more widely). First, we posit a family of factorizable probability distributions on latent variables

$$\mathcal{Q}^{n+d} = \left\{ q : q(\theta, z) = \prod_{i=1}^{d} q_{\theta_i}(\theta_i) \prod_{j=1}^{n} q_{z_j}(z_j) \right\}.$$

This family is called *the mean-field family*. It represents a joint of the latent variables with $n + d$ (parametric) marginal distributions, $\{q_{\theta_1}, \dots, q_{\theta_d}, q_{z_1}, \dots, q_{z_n}\}$.

VB finds the member of the family closest to the exact posterior $p(\theta, z \mid x)$, where closeness is measured by KL divergence. Thus VB seeks to solve the optimization,

$$q^*(\theta, z) = \underset{q(\theta,z)\in\mathcal{Q}^{n+d}}{\arg\min} \ \mathrm{KL}(q(\theta, z) \,||\, p(\theta, z|x)). \qquad (2)$$

In practice, VB finds $q^*(\theta, z)$ by optimizing an alternative objective, *the evidence lower bound (ELBO)*,

$$\mathrm{ELBO}(q(\theta, z)) = -\int q(\theta, z) \log \frac{q(\theta, z)}{p(\theta, x, z)} d\theta dz. \qquad (3)$$

This objective is called the ELBO because it is a lower bound on the evidence $\log p(x)$. More importantly, the ELBO is equal to the negative KL plus $\log p(x)$, which does not depend on $q(\cdot)$. Maximizing the ELBO minimizes the KL (Jordan et al. 1999).

The optimum $q^*(\theta, z) = q^*(\theta)q^*(z)$ approximates the posterior, and we call it *the* VB *posterior.* (For simplicity, we will write $q(\theta, z) = \prod_{i=1}^{d} q(\theta_i) \prod_{j=1}^{n} q(z_j)$, omitting the subscript on the factors $q(\cdot)$. The understanding is that the factor is indicated by its argument.) Though it cannot capture posterior dependence across latent variables, it has hope to capture each of their marginals. In particular, this article is about the theoretical properties of the VB posterior $q^*(\theta)$, the VB posterior of θ. We will also focus on the corresponding expectation of the global variable, that is, an estimate of the parameter. It is

$$\hat{\theta}_n^* := \int \theta \cdot q^*(\theta) \mathrm{d}\theta.$$

We call θ^* *the variational Bayes estimate* (VBE).

VB methods are fast and yield good predictive performance in empirical experiments (Blei, Kucukelbir, and McAuliffe 2016). However, there are few rigorous theoretical results. In this article, we prove that (1) the VB posterior converges in total variation (TV) distance to the KL minimizer of a normal distribution centered at the truth and (2) the VBE is consistent and asymptotically normal.

These theorems are frequentist in the sense that we assume the data come from $p(x, z\,;\, \theta_0)$ with a true (nonrandom) θ_0. We then study properties of the corresponding posterior distribution $p(\theta \mid x)$, when approximating it with variational inference. What this work shows is that the VB posterior is consistent even though the mean field approximating family can be a brutal approximation. In this sense, VB is a theoretically sound approximate inference procedure.

1.1. Main Ideas

We describe the results of the article. Along the way, we will need to define some terms: the variational frequentist estimate (VFE), the variational log-likelihood, the VB posterior, the VBE, and the VB ideal. Our results center around the VB posterior and the VBE. (Table 1 contains a glossary of terms.)

Table 1. Glossary of terms.

Name	Definition
Variational log-likelihood	$M_n(\theta\,;\,x) := \sup_{q(z)\in\mathcal{Q}^n} \int q(z) \log \frac{p(x,z\|\theta)}{q(z)} dz$
Variational frequentist estimate (VFE)	$\hat{\theta}_n := \arg\max_\theta M_n(\theta\,;\,x)$
VB ideal	$\pi^*(\theta\|x) := \frac{p(\theta)\exp\{M_n(\theta\,;\,x)\}}{\int p(\theta)\exp\{M_n(\theta\,;\,x)\}d\theta}$
Evidence Lower Bound (ELBO)	$\mathrm{ELBO}(q(\theta, z)) := \int\int q(\theta)q(z) \log \frac{p(x,z,\theta)}{q(\theta)q(z)} d\theta dz$
VB posterior	$q^*(\theta) := \arg\max_{q(\theta)\in\mathcal{Q}^d} \sup_{q(z)\in\mathcal{Q}^n} ELBO(q(\theta, z))$
VB estimate (VBE)	$\hat{\theta}_n^* := \int \theta \cdot q^*(\theta) d\theta$

The variational frequentist estimate (VFE) and the variational log-likelihood. The first idea that we define is the *variational frequentist estimate* (VFE). It is a point estimate of θ that maximizes a local variational objective with respect to an optimal variational distribution of the local variables. (The VFE treats the variable θ as a parameter rather than a random variable.) We call the objective *the variational log-likelihood*,

$$M_n(\theta\,;\,x) = \max_{q(z)} \ \mathbb{E}_{q(z)}\left[\log p(x, z \mid \theta) - \log q(z)\right]. \qquad (4)$$

In this objective, the optimal variational distribution $q^\dagger(z)$ solves the local variational inference problem,

$$q^\dagger(z) = \arg\min_q \ \mathrm{KL}(q(z) \,||\, p(z \mid x, \theta)). \qquad (5)$$

Note that $q^\dagger(z)$ implicitly depends on both the data x and the parameter θ.

With the objective defined, the VFE is

$$\hat{\theta}_n = \arg\max_\theta \ M_n(\theta\,;\,x). \qquad (6)$$

It is usually calculated with variational expectation maximization (EM) (Wainwright and Jordan 2008; Ormerod and Wand 2010), which iterates between the E step of Equation (5) and the M step of Equation (6). Recent research has explored the theoretical properties of the VFE for stochastic block models (Bickel et al. 2013), generalized linear mixed models (Hall et al. 2011b), and Gaussian mixture models (Westling and McCormick 2015).

We make two remarks. First, the maximizing variational distribution $q^\dagger(z)$ of Equation (5) is different from $q^*(z)$ in the VB posterior: $q^\dagger(z)$ is implicitly a function of individual values of θ, while $q^*(z)$ is implicitly a function of the variational distributions $q(\theta)$. Second, the variational log-likelihood in Equation (4) is similar to the original objective function for the EM algorithm (Dempster, Laird, and Rubin 1977). The difference is that the EM objective is an expectation with respect to the exact conditional $p(z \mid x)$, whereas the variational log-likelihood uses a variational distribution $q(z)$.

Variational Bayes and ideal variational Bayes. While earlier applications of variational inference appealed to variational EM and the VFE, most modern applications do not. Rather they use VB, as we described above, where there is a prior on θ and we approximate its posterior with a global variational distribution $q(\theta)$. One advantage of VB is that it provides regularization through the prior. Another is that it requires only one type of optimization: the same considerations around updating the local variational factors $q(z)$ are also at play when updating the global factor $q(\theta)$.

To develop theoretical properties of VB, we connect the VB posterior to the variational log-likelihood; this is a stepping stone to the final analysis. In particular, we define the VB *ideal posterior* $\pi^*(\theta \mid x)$,

$$\pi^*(\theta \mid x) = \frac{p(\theta)\exp\{M_n(\theta\,;\,x)\}}{\int p(\theta)\exp\{M_n(\theta\,;\,x)\}d\theta}. \tag{7}$$

Here, the local latent variables z are constrained under the variational family but the global latent variables θ are not. Note that because it depends on the variational log-likelihood $M_n(\theta\,;\,x)$, this distribution implicitly contains an optimal variational distribution $q^\dagger(z)$ for each value of θ; see Equations (4) and (5).

Loosely, the VB ideal lies between the exact posterior $p(\theta \mid x)$ and a variational approximation $q(\theta)$. It recovers the exact posterior when $p(z \mid \theta, x)$ degenerates to a point mass and $q^\dagger(z)$ is always equal to $p(z \mid \theta, x)$; in that case the variational likelihood is equal to the log-likelihood and Equation (7) is the posterior. But $q^\dagger(z)$ is usually an approximation to the conditional. Thus the VB ideal usually falls short of the exact posterior.

That said, the VB ideal is more complex that a simple parametric variational factor $q(\theta)$. The reason is that its value for each θ is defined by the optimization within $M_n(\theta\,;\,x)$. Such a distribution will usually lie outside the distributions attainable with a simple family.

In this work, we first establish the theoretical properties of the VB ideal. We then connect it to the VB posterior.

Variational Bernstein–von Mises. We have set up the main concepts. We now describe the main results.

Suppose the data come from a true (finite-dimensional) parameter θ_0. The classical Bernstein–von Mises theorem says that, under certain conditions, the exact posterior $p(\theta \mid x)$ approaches a normal distribution, independent of the prior, as the number of observations tends to infinity. In this article, we extend the theory around Bernstein–von Mises to the variational posterior. Here we summarize our results.

- Lemma 1 shows that the VB ideal $\pi^*(\theta \mid x)$ is consistent and converges to a normal distribution around the VFE. If the VFE is consistent, the VB ideal $\pi^*(\theta|x)$ converges to a normal distribution whose mean parameter is a random vector centered at the true parameter. (Note the randomness in the mean parameter is due to the randomness in the observations x.)
- We next consider the point in the variational family that is closest to the VB ideal $\pi^*(\theta|x)$ in KL divergence. Lemma 2 and Lemma 3 show that this KL minimizer is consistent and converges to the KL minimizer of a normal distribution around the VFE. If the VFE is consistent (Hall et al. 2011b; Bickel et al. 2013), then the KL minimizer converges to the KL minimizer of a normal distribution with a random mean centered at the true parameter.
- Lemma 4 shows that the VB posterior $q^*(\theta)$ enjoys the same asymptotic properties as the KL minimizers of the VB ideal $\pi^*(\theta \mid x)$.
- Theorem 5 is the *variational Bernstein–von Mises theorem*. It shows that the VB posterior $q^*(\theta)$ is asymptotically normal around the VFE. Again, if the VFE is consistent then the VB posterior converges to a normal with a random mean centered at the true parameter. Further, Theorem 6 shows that the VBE $\hat{\theta}_n^*$ is consistent with the true parameter and asymptotically normal.
- Finally, we prove two corollaries. First, if we use a full-rank Gaussian variational family, then the corresponding VB posterior recovers the true mean and covariance. Second, if we use a mean-field Gaussian variational family, then the VB posterior recovers the true mean and the marginal variance, but not the off-diagonal terms. The mean-field VB posterior is underdispersed.

Related work. This work draws on two themes. The first is the body of work on theoretical properties of variational inference. You, Ormerod, and Müller (2014) and Ormerod, You, and Muller (2014) studied variational Bayes for a classical Bayesian linear model. They used normal priors and spike-and-slab priors on the coefficients, respectively. Wang and Titterington (2004) studied variational Bayesian approximations for exponential family models with missing values. Wang and Titterington (2005) and Wang and Titterington (2006) analyzed variational Bayes in Bayesian mixture models with conjugate priors. More recently, Zhang and Zhou (2017) studied mean field variational inference in stochastic block model (SBMs) with a batch coordinate ascent algorithm: it has a linear convergence rate and converges to the minimax rate within $\log n$ iterations. Sheth and Khardon (2017) proved a bound for the excess Bayes risk using variational inference in latent Gaussian models. Ghorbani, Javadi, and Montanari (2018) studied a version of latent Dirichlet allocation (LDA) and identified an instability in variational inference in certain signal-to-noise ratio (SNR) regimes. Zhang and Gao (2017) characterized the convergence rate of variational posteriors for nonparametric and high-dimensional inference. Pati, Bhattacharya, and Yang (2017) provided general conditions for obtaining optimal risk bounds for point estimates acquired from mean field variational Bayes. Alquier, Ridgway, and Chopin (2016) and Alquier and Ridgway (2017) studied the concentration of variational approximations of Gibbs posteriors and fractional posteriors based on PAC-Bayesian inequalities. Yang, Pati, and Bhattacharya (2017) proposed α-variational inference and developed variational inequalities for the Bayes risk under the variational solution.

On the frequentist side, Hall, Ormerod, and Wand (2011a); Hall et al. (2011b) established the consistency of Gaussian variational EM estimates in a Poisson mixed-effects model with a single predictor and a grouped random intercept. Westling and McCormick (2015) studied the consistency of variational EM estimates in mixture models through a connection to M-estimation. Celisse, Daudin, and Pierre (2012) and Bickel et al. (2013) proved the asymptotic normality of parameter estimates in the SBM under a mean field variational approximation.

However, many of these treatments of variational methods—Bayesian or frequentist—are constrained to specific models and priors. Our work broadens these works by considering more general models. Moreover, the frequentist works focus on estimation procedures under a variational approximation. We expand on these works by proving a variational Bernstein–von

Mises theorem, leveraging the frequentist results to analyze VB posteriors.

The second theme is the Bernstein–von Mises theorem. The classical (parametric) Bernstein–von Mises theorem roughly says that the posterior distribution of $\sqrt{n}(\theta - \theta_0)$ "converges," under the true parameter value θ_0, to $\mathcal{N}(X, 1/I(\theta_0))$, where $X \sim \mathcal{N}(0, 1/I(\theta_0))$ and $I(\theta_0)$ is the Fisher information (Le Cam 1953; Van der Vaart 2000; Ghosh and Ramamoorthi 2003; Le Cam and Yang 2012). Early forms of this theorem date back to Laplace, Bernstein, and von Mises (Laplace 1809; Bernstein 1917; Von Mises 1931). A version also appeared in Lehmann and Casella (2006). Kleijn and Van der Vaart (2012) established the Bernstein–von Mises theorem under model misspecification. Recent advances include extensions to semiparametric cases (Murphy and Van der Vaart 2000; Kim 2006; De Blasi and Hjort 2009; Rivoirard and Rousseau 2012; Bickel and Kleijn 2012; Castillo 2012a, 2012b; Castillo and Nickl 2014; Panov and Spokoiny 2014; Castillo and Rousseau 2015; Ghosal and van der Vaart 2017) and nonparametric cases (Diaconis and Freedman 1986; Cox 1993; Diaconis and Freedman 1997; 1998; Freedman 1999; Kim and Lee 2004; James 2008; Boucheron et al. 2009; Kim 2009; Johnstone 2010; Bontemps et al. 2011; Knapik et al. 2011; Leahu 2011; Rivoirard et al. 2012; Castillo and Nickl 2012; 2013; Spokoiny 2013; Castillo 2012, 2014; Ray 2017; Panov and Spokoiny 2015; Lu 2017). In particular, Lu, Stuart, and Weber (2016) proved a Bernstein–von Mises type result for Bayesian inverse problems, characterizing Gaussian approximations of probability measures with respect to the KL divergence. Below, we borrow proof techniques from Lu, Stuart, and Weber (2016). But we move beyond the Gaussian approximation to establish the consistency of variational Bayes.

This article. The rest of the article is organized as follows. Section 2 characterizes theoretical properties of the VB ideal. Section 3 contains the central results of the article. It first connects the VB ideal and the VB posterior. It then proves the variational Bernstein–von Mises theorem, which characterizes the asymptotic properties of the VB posterior and VB estimate. Section 4 studies three models under this theoretical lens, illustrating how to establish consistency and asymptotic normality of specific VB estimates. Section 5 reports simulation studies to illustrate these theoretical results. Finally, Section 6 concludes with article with a discussion.

2. The VB Ideal

To study the VB posterior $q^*(\theta)$, we first study the VB ideal of Equation (7). In the next section, we connect it to the VB posterior.

Recall the VB ideal is

$$\pi^*(\theta|x) = \frac{p(\theta)\exp(M_n(\theta\,;\,x))}{\int p(\theta)\exp(M_n(\theta\,;\,x))d\theta},$$

where $M_n(\theta\,;\,x)$ is the variational log-likelihood of Equation (4). If we embed the variational log-likelihood $M_n(\theta\,;\,x)$ in a statistical model of x, this model has likelihood

$$\ell(\theta\,;\,x) \propto \exp(M_n(\theta; x)).$$

We call it the *frequentist variational model*. The VB ideal $\pi^*(\theta|x)$ is thus the classical posterior under the frequentist variational model $\ell(\theta\,;\,x)$; the VFE is the classical maximum likelihood estimate (MLE).

Consider the results around frequentist estimation of θ under variational approximations of the local variables z (Hall et al. 2011b; Bickel et al. 2013; Westling and McCormick 2015). These works consider asymptotic properties of estimators that maximize $M_n(\theta\,;\,x)$ with respect to θ. We will first leverage these results to prove properties of the VB ideal and their KL minimizers in the mean field variational family $\mathcal{Q}^d$. Then we will use these properties to study the VB posterior, which is what is estimated in practice.

This section relies on the consistent testability and the local asymptotic normality (LAN) of $M_n(\theta\,;\,x)$ (defined later) to show the VB ideal is consistent and asymptotically normal. We will then show that its KL minimizer in the mean field family is also consistent and converges to the KL minimizer of a normal distribution in TV distance.

These results are not surprising. Suppose the variational log-likelihood behaves similarly to the true log-likelihood, that is, they produce consistent parameter estimates. Then, in the spirit of the classical Bernstein–von Mises theorem under model misspecification (Kleijn et al. 2012), we expect the VB ideal to be consistent as well. Moreover, the approximation through a factorizable variational family should not ruin this consistency—point masses are factorizable and thus the limiting distribution lies in the approximating family.

2.1. The VB Ideal

The lemma statements and proofs adapt ideas from Ghosh and Ramamoorthi (2003); Van der Vaart (2000); Bickel and Yahav (1967); Kleijn et al. (2012); Lu, Stuart, and Weber (2017) to the variational log-likelihood. Let Θ be an open subset of $\mathbb{R}^d$. Suppose the observations $x = x_{1:n}$ are a random sample from the measure P_{θ_0} with density $\int p(x, z|\theta = \theta_0)dz$ for some fixed, nonrandom value $\theta_0 \in \Theta$. $z = z_{1:n}$ are local latent variables, and $\theta = \theta_{1:d} \in \Theta$ are global latent variables. We assume that the density maps $(\theta, x) \mapsto \int p(x, z|\theta)dz$ of the true model and $(\theta, x) \mapsto \ell(\theta\,;\,x)$ of the variational frequentist models are measurable. For simplicity, we also assume that for each n there exists a single measure that dominates all measures with densities $\ell(\theta\,;\,x), \theta \in \Theta$ as well as the true measure P_{θ_0}.

Assumption 1. We assume the following conditions for the rest of the article:

1. (Prior mass) The prior measure with Lebesgue-density $p(\theta)$ on Θ is continuous and positive on a neighborhood of θ_0. There exists a constant $M_p > 0$ such that $|(\log p(\theta))''| \leq M_p e^{|\theta|^2}$.
2. (Consistent testability) For every $\epsilon > 0$, there exists a sequence of tests ϕ_n such that

$$\int \phi_n(x)p(x, z|\theta_0)dzdx \to 0$$

and

$$\sup_{\theta:||\theta-\theta_0||\geq\epsilon} \int (1 - \phi_n(x))\frac{\ell(\theta\,;\,x)}{\ell(\theta_0\,;\,x)}p(x, z|\theta_0)dzdx \to 0,$$

3. (Local asymptotic normality (LAN)) For every compact set $K \subset \mathbb{R}^d$, there exist random vectors Δ_{n,θ_0} bounded in probability and nonsingular matrices V_{θ_0} such that

$$\sup_{h \in K} |M_n(\theta + \delta_n h\,;\, x) \ - M_n(\theta\,;\, x) - h^\top V_{\theta_0} \Delta_{n,\theta_0} + \frac{1}{2} h^\top V_{\theta_0} h| \xrightarrow{P_{\theta_0}} 0,$$

where δ_n is a $d \times d$ diagonal matrix. We have $\delta_n \to 0$ as $n \to \infty$. For $d = 1$, we commonly have $\delta_n = 1/\sqrt{n}$.

These three assumptions are standard for Bernstein–von Mises theorem. The first assumption is a prior mass assumption. It says the prior on θ puts enough mass to sufficiently small balls around θ_0. This allows for optimal rates of convergence of the posterior. The first assumption further bounds the second derivative of the log prior density. This is a mild technical assumption satisfied by most nonheavy-tailed distributions.

The second assumption is a consistent testability assumption. It says there exists a sequence of uniformly consistent (under P_{θ_0}) tests for testing $H_0 : \theta = \theta_0$ against $H_1 : ||\theta - \theta_0|| \geq \epsilon$ for every $\epsilon > 0$ based on the frequentist variational model. This is a weak assumption. For example, it suffices to have a compact Θ and continuous and identifiable $M_n(\theta\,;\, x)$. It is also true when there exists a consistent estimator T_n of θ. In this case, we can set $\phi_n := 1\{T_n - \theta \geq \epsilon/2\}$.

The last assumption is a local asymptotic normality assumption on $M_n(\theta\,;\, x)$ around the true value θ_0. It says the frequentist variational model can be asymptotically approximated by a normal location model centered at θ_0 after a rescaling of δ_n^{-1}. This normalizing sequence δ_n determines the optimal rates of convergence of the posterior. For example, if $\delta_n = 1/\sqrt{n}$, then we commonly have $\theta - \theta_0 = O_p(1/\sqrt{n})$. We often need model-specific analysis to verify this condition, as we do in Section 4. We discuss sufficient conditions and general proof strategies in Section 3.4.

In the spirit of the last assumption, we perform a change-of-variable step:

$$\tilde{\theta} = \delta_n^{-1}(\theta - \theta_0). \tag{8}$$

We center θ at the true value θ_0 and rescale it by the reciprocal of the rate of convergence δ_n^{-1}. This ensures that the asymptotic distribution of $\tilde{\theta}$ is not degenerate, that is, it does not converge to a point mass. We define $\pi^*_{\tilde{\theta}}(\cdot|x)$ as the density of $\tilde{\theta}$ when θ has density $\pi^*(\cdot|x)$:

$$\pi^*_{\tilde{\theta}}(\tilde{\theta}|x) = \pi^*(\theta_0 + \delta_n \tilde{\theta}|x) \cdot |\det(\delta_n)|.$$

Now we characterize the asymptotic properties of the VB ideal.

Lemma 1. The VB ideal converges in total variation to a sequence of normal distributions,

$$||\pi^*_{\tilde{\theta}}(\cdot|x) - \mathcal{N}\left(\cdot\,;\, \Delta_{n,\theta_0}, V_{\theta_0}^{-1}\right) ||_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0.$$

Proof sketch of Lemma 1. This is a consequence of the classical finite-dimensional Bernstein–von Mises theorem under model misspecification (Kleijn et al. 2012). Theorem 2.1 of Kleijn and Van der Vaart (2012) roughly says that the posterior is consistent if the model is locally asymptotically normal around the true parameter value θ_0. Here, the true data-generating measure is P_{θ_0} with density $\int p(x, z|\theta = \theta_0)dz$, while the frequentist variational model has densities $\ell(\theta\,;\, x), \theta \in \Theta$.

What we need to show is that the consistent testability assumption in Assumption 1 implies assumption (2.3) in Kleijn et al. (2012):

$$\int_{|\tilde{\theta}|>M_n} \pi^*_{\tilde{\theta}}(\tilde{\theta}|x) d\tilde{\theta} \xrightarrow{P_{\theta_0}} 0$$

for every sequence of constants $M_n \to \infty$. To show this, we mimic the argument of Theorem 3.1 of Kleijn et al. (2012), where they show this implication for the iid case with a common convergence rate for all dimensions of θ. See Appendix A for details. □

This lemma says the VB ideal of the rescaled θ, $\tilde{\theta} = \delta_n^{-1}(\theta - \theta_0)$, is asymptotically normal with mean Δ_{n,θ_0}. The mean, Δ_{n,θ_0}, as assumed in Assumption 1, is a random vector bounded in probability. The asymptotic distribution $\mathcal{N}(\cdot\,;\, \Delta_{n,\theta_0}, V_{\theta_0}^{-1})$ is thus also random, where randomness is due to the data x being random draws from the true data-generating measure $\mathbb{P}_{\theta_0}$. We notice that if the VFE, $\hat{\theta}_n$, is consistent and asymptotically normal, we commonly have $\Delta_{n,\theta_0} = \delta_n^{-1}(\hat{\theta}_n - \theta_0)$ with $\mathbb{E}(\Delta_{n,\theta_0}) = 0$. Hence, the VB ideal will converge to a normal distribution with a random mean centered at the true value θ_0.

2.2. The KL Minimizer of the VB Ideal

Next we study the KL minimizer of the VB ideal in the mean field variational family. We show its consistency and asymptotic normality. To be clear, the asymptotic normality is in the sense that the KL minimizer of the VB ideal converges to the KL minimizer of a normal distribution in TV distance.

Lemma 2. The KL minimizer of the VB ideal over the mean field family is consistent: almost surely under P_{θ_0}, it converges to a point mass,

$$\operatorname*{arg\,min}_{q(\theta) \in \mathcal{Q}^d} \mathrm{KL}(q(\theta)||\pi^*(\theta|x)) \xrightarrow{d} \delta_{\theta_0}.$$

Proof sketch of Lemma 2. The key insight here is that point masses are factorizable. Lemma 1 above suggests that the VB ideal converges in distribution to a point mass. We thus have its KL minimizer also converging to a point mass, because point masses reside within the mean field family. In other words, there is no loss, in the limit, incurred by positing a factorizable variational family for approximation. □

To prove this lemma, we bound the mass of $B^c(\theta_0, \eta_n)$ under $q(\theta)$, where $B^c(\theta_0, \eta_n)$ is the complement of an η_n-sized ball centered at θ_0 with $\eta_n \to 0$ as $n \to \infty$. In this step, we borrow ideas from the proof of Lemma 3.6 and Lemma 3.7 in Lu, Stuart, and Weber (2017). See Appendix B for details.

Lemma 3. The KL minimizer of the VB ideal of $\tilde{\theta}$ converges to that of $\mathcal{N}(\cdot\,;\, \Delta_{n,\theta_0}, V_{\theta_0}^{-1})$ in total variation: under mild technical conditions on the tail behavior of $\mathcal{Q}^d$ (see Assumption 2 in

Appendix C),

$$\Bigg\| \mathop{\arg\min}_{q\in\mathcal{Q}^d} \mathrm{KL}(q(\cdot)||\pi^*_{\tilde\theta}(\cdot|x)) - \mathop{\arg\min}_{q\in\mathcal{Q}^d} \mathrm{KL}\left(q(\cdot)||\mathcal{N}\left(\cdot\,;\ \Delta_{n,\theta_0}, V_{\theta_0}^{-1}\right)\right)\Bigg\|_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0.$$

Proof sketch of Lemma 3. The intuition here is that, if the two distribution are close in the limit, their KL minimizers should also be close in the limit. Lemma 1 says that the VB ideal of $\tilde\theta$ converges to $\mathcal{N}(\cdot;\Delta_{n,\theta_0}, V_{\theta_0}^{-1})$ in total variation. We would expect their KL minimizer also converges in some metric. This result is also true for the (full-rank) Gaussian variational family if rescaled appropriately.

Here we show their convergence in total variation. This is achieved by showing the Γ-convergence of the functionals of q: $\mathrm{KL}(q(\cdot)||\pi^*_{\tilde\theta}(\cdot|x))$ to $\mathrm{KL}(q(\cdot)||\mathcal{N}(\cdot\,;\ \Delta_{n,\theta_0}, V_{\theta_0}^{-1}))$, for parametric q's. Γ-convergence is a classical tool for characterizing variational problems; Γ-convergence of functionals ensures convergence of their minimizers (Braides 2006; Dal Maso 2012). See Appendix C for proof details and a review of Γ-convergence. □

We characterized the limiting properties of the VB ideal and their KL minimizers. We will next show that the VB posterior is close to the KL divergence minimizer of the VB ideal. Section 3 culminates in the main theorem of this article—the variational Bernstein–von Mises theorem—showing the VB posterior share consistency and asymptotic normality with the KL divergence minimizer of VB ideal.

3. Frequentist Consistency of Variational Bayes

We now study the VB posterior. In the previous section, we proved theoretical properties for the VB ideal and its KL minimizer in the variational family. Here, we first connect the VB ideal to the VB posterior, the quantity that is used in practice. We then use this connection to understand the theoretical properties of the VB posterior.

We begin by characterizing the optimal variational distribution in a useful way. Decompose the variational family as

$$q(\theta, z) = q(\theta)q(z),$$

where $q(\theta) = \prod_{i=1}^d q(\theta_i)$ and $q(z) = \prod_{i=1}^n q(z_i)$. Denote the prior $p(\theta)$. Note d does not grow with the size of the data. We will develop a theory around VB that considers asymptotic properties of the VB posterior $q^*(\theta)$.

We decompose the ELBO of Equation (3) into the portion associated with the global variable and the portion associated with the local variables,

$$\begin{aligned}\mathrm{ELBO}(q(\theta)q(z)) &= \int\int q(\theta)q(z)\log\frac{p(\theta,x,z)}{q(\theta)q(z)}d\theta dz\\ &= \int\int q(\theta)q(z)\log\frac{p(\theta)p(x,z|\theta)}{q(\theta)q(z)}d\theta dz\\ &= \int q(\theta)\log\frac{p(\theta)}{q(\theta)}d\theta\\ &\quad + \int q(\theta)\int q(z)\log\frac{p(x,z\,|\,\theta)}{q(z)}d\theta dz.\end{aligned}$$

The optimal variational factor for the global variables, that is, the VB posterior, maximizes the ELBO. From the decomposition, we can write it as a function of the optimized local variational factor,

$$\begin{aligned}q^*(\theta) = \mathop{\arg\max}_{q(\theta)}\ \sup_{q(z)} \int q(\theta)\Bigg(\log\Bigg[p(\theta)\\ \times\exp\left\{\int q(z)\log\frac{p(x,z|\theta)}{q(z)}dz\right\}\Bigg] - \log q(\theta)\Bigg)d\theta. \quad (9)\end{aligned}$$

One way to see the objective for the VB posterior is as the ELBO profiled over $q(z)$, that is, where the optimal $q(z)$ is a function of $q(\theta)$ (Hoffman et al. 2013). With this perspective, the ELBO becomes a function of $q(\theta)$ only. We denote it as a functional $\mathrm{ELBO}_p(\cdot)$:

$$\begin{aligned}\mathrm{ELBO}_p(q(\theta)) := \sup_{q(z)} \int q(\theta)\Bigg(\log\Bigg[p(\theta)\\ \times\exp\left\{\int q(z)\log\frac{p(x,z|\theta)}{q(z)}dz\right\}\Bigg] - \log q(\theta)\Bigg)d\theta.\end{aligned} \quad (10)$$

We then rewrite Equation (9) as $q^*(\theta) = \arg\max_{q(\theta)} \mathrm{ELBO}_p(q(\theta))$. This expression for the VB posterior is key to our results.

3.1. KL Minimizers of the VB Ideal

Recall that the KL minimization objective to the ideal VB posterior is the functional $\mathrm{KL}(\cdot||\pi^*(\theta|x))$. We first show that the two optimization objectives $\mathrm{KL}(\cdot||\pi^*(\theta|x))$ and $\mathrm{ELBO}_p(\cdot)$ are close in the limit. Given the continuity of both $\mathrm{KL}(\cdot||\pi^*(\theta|x))$ and $\mathrm{ELBO}_p(\cdot)$, this implies the asymptotic properties of optimizers of $\mathrm{KL}(\cdot||\pi^*(\theta|x))$ will be shared by the optimizers of $\mathrm{ELBO}_p(\cdot)$.

Lemma 4. The negative KL divergence to the VB ideal is equivalent to the profiled ELBO in the limit: under mild technical conditions on the tail behavior of $\mathcal{Q}^d$ (see, e.g., Assumption 3 in Appendix D), for $q(\theta)\in\mathcal{Q}^d$,

$$\mathrm{ELBO}_p(q(\theta)) = -\mathrm{KL}(q(\theta)||\pi^*(\theta|x)) + o_P(1).$$

Proof sketch of Lemma 4. We first notice that

$$\begin{aligned}&-\mathrm{KL}(q(\theta)||\pi^*(\theta|x)) && (11)\\ &= \int q(\theta)\log\frac{p(\theta)\exp(M_n(\theta\,;\ x))}{q(\theta)}d\theta && (12)\\ &= \int q(\theta)\Bigg(\log\Bigg[p(\theta)\\ &\quad\times\exp\left\{\sup_{q(z)}\int q(z)\log\frac{p(x,z|\theta)}{q(z)}dz\right\}\Bigg] - \log q(\theta)\Bigg)d\theta. && (13)\end{aligned}$$

Comparing Equation (13) with Equation (10), we can see that the only difference between $-\mathrm{KL}(\cdot||\pi^*(\theta|x))$ and $\mathrm{ELBO}_p(\cdot)$ is in the position of $\sup_{q(z)}$. $\mathrm{ELBO}_p(\cdot)$ allows for a single choice of optimal $q(z)$ given $q(\theta)$, while $-\mathrm{KL}(\cdot||\pi^*(\theta|x))$ allows for a different optimal $q(z)$ for each value of θ. In this sense, if we restrict the variational family of $q(\theta)$ to be point masses, then $\mathrm{ELBO}_p(\cdot)$ and $-\mathrm{KL}(\cdot||\pi^*(\theta|x))$ will be the same.

The only members of the variational family of $q(\theta)$ that admit finite $-\mathrm{KL}(q(\theta)||\pi^*(\theta|x))$ are ones that converge to point masses at rate δ_n, so we expect $\mathrm{ELBO}_p(\cdot)$ and $-\mathrm{KL}(\cdot||\pi^*(\theta|x))$ to be close as $n \to \infty$. We prove this by bounding the remainder in the Taylor expansion of $M_n(\theta\,;\,x)$ by a sequence converging to zero in probability. See Appendix D for details. □

3.2. The VB Posterior

Section 2 characterizes the asymptotic behavior of the VB ideal $\pi^*(\theta|x)$ and their KL minimizers. Lemma 4 establishes the connection between the VB posterior $q^*(\theta)$ and the KL minimizers of the VB ideal $\pi^*(\theta|x)$. Recall $\arg\min_{q(\theta)\in\mathcal{Q}^d} \mathrm{KL}(q(\theta)||\pi^*(\theta|x))$ is consistent and converges to the KL minimizer of a normal distribution. We now build on these results to study the VB posterior $q^*(\theta)$.

Now we are ready to state the main theorem. It establishes the asymptotic behavior of the VB posterior $q^*(\theta)$.

Theorem 5 (Variational Bernstein–von-Mises Theorem).

1. The VB posterior is consistent: almost surely under P_{θ_0},

$$q^*(\theta) \xrightarrow{d} \delta_{\theta_0}.$$

2. The VB posterior is asymptotically normal in the sense that it converges to the KL minimizer of a normal distribution:

$$\left\| q^*_{\tilde{\theta}}(\cdot) - \arg\min_{q\in\mathcal{Q}^d} \mathrm{KL}\left(q(\cdot)||\mathcal{N}\left(\cdot\,;\,\Delta_{n,\theta_0}, V_{\theta_0}^{-1}\right)\right)\right\|_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0. \tag{14}$$

Here we transform $q^*(\theta)$ to $q_{\tilde{\theta}}(\tilde{\theta})$, which is centered around the true θ_0 and scaled by the convergence rate; see Equation (8). When $\mathcal{Q}^d$ is the mean field variational family, then the limiting VB posterior is normal:

$$\arg\min_{q\in\mathcal{Q}^d} \mathrm{KL}\left(q(\cdot)||\mathcal{N}\left(\cdot\,;\,\Delta_{n,\theta_0}, V_{\theta_0}^{-1}\right)\right) = \mathcal{N}(\cdot\,;\,\Delta_{n,\theta_0}, V_{\theta_0}^{\prime -1})), \tag{15}$$

where V'_{θ_0} is diagonal and has the same diagonal terms as V_{θ_0}.

Proof sketch of Theorem 5. This theorem is a direct consequence of Lemma 2, Lemma 3, Lemma 4. We need the same mild technical conditions on $\mathcal{Q}^d$ as in Lemma 3 and Lemma 4. Equation (15) can be proved by first establishing the normality of the optimal variational factor (see Section 10.1.2 of Bishop 2006 for details) and proceeding with Lemma 8. See Appendix E for details. □

Given the convergence of the VB posterior, we can now establish the asymptotic properties of the VBE.

Theorem 6 (Asymptotics of the VBE). Assume $\int |\theta|^2 \pi(\theta)\,d\theta < \infty$. Let $\hat{\theta}^*_n = \int \theta \cdot q^*_1(\theta)\,d\theta$ denote the VBE.

1. The VBE is consistent: under $\mathbb{P}_{\theta_0}$,

$$\hat{\theta}^*_n \xrightarrow{a.s.} \theta_0.$$

2. The VBE is asymptotically normal in the sense that it converges in distribution to the mean of the KL minimizer (The randomness in the mean of the KL minimizer comes from Δ_{n,θ_0}): if $\Delta_{n,\theta_0} \xrightarrow{d} X$ for some X,

$$\delta_n^{-1}(\hat{\theta}^*_n - \theta_0) \xrightarrow{d} \int \tilde{\theta} \cdot \arg\min_{q\in\mathcal{Q}^d} \mathrm{KL}\left(q(\tilde{\theta})||\mathcal{N}\left(\tilde{\theta}\,;\,X, V_{\theta_0}^{-1}\right)\right) d\tilde{\theta}.$$

Proof sketch of Theorem 6. As the posterior mean is a continuous function of the posterior distribution, we would expect the VBE is consistent given the VB posterior is. We also know that the posterior mean is the Bayes estimator under squared loss. Thus, we would expect the VBE to converge in distribution to squared loss minimizer of the KL minimizer of the VB ideal. The result follows from a very similar argument from Theorem 2.3 of Kleijn and Van der Vaart (2012), which shows that the posterior mean estimate is consistent and asymptotically normal under model misspecification as a consequence of the Bernstein–von Mises theorem and the argmax theorem. See Appendix E for details. □

We remark that Δ_{n,θ_0}, as in Assumption 1, is a random vector bounded in P_{θ_0} probability. The randomness is due to x being a random sample generated from P_{θ_0}. In cases where VFE is consistent, like in all the examples we will see in Section 4, Δ_{n,θ_0} is a zero mean random vector with finite variance. For particular realizations of x the value of Δ_{n,θ_0} might not be zero; however, because we scale by δ_n^{-1}, this does not destroy the consistency of VB posterior or the VBE.

3.3. Gaussian VB Posteriors

We illustrate the implications of Theorem 5 and Theorem 6 on two choices of variational families: a full-rank Gaussian variational family and a factorizable Gaussian variational family. In both cases, the VB posterior and the VBE are consistent and asymptotically normal with different covariance matrices. The VB posterior under the factorizable family is underdispersed.

Corollary 7. Posit a full-rank Gaussian variational family, that is,

$$\mathcal{Q}^d = \{q : q(\theta) = \mathcal{N}(m, \Sigma)\}, \tag{16}$$

with Σ positive definite. Then

1. $q^*(\theta) \xrightarrow{d} \delta_{\theta_0}$, almost surely under $\mathbb{P}_{\theta_0}$.
2. $||q^*_{\tilde{\theta}}(\cdot) - \mathcal{N}(\cdot\,;\,\Delta_{n,\theta_0}, V_{\theta_0}^{-1})||_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0.$
3. $\hat{\theta}^*_n \xrightarrow{a.s.} \theta_0.$
4. $\delta_n^{-1}(\hat{\theta}^*_n - \theta_0) - \Delta_{n,\theta_0} = o_{P_{\theta_0}}(1).$

Proof sketch of Corollary 7. This is a direct consequence of Theorem 5 and Theorem 6. We only need to show that Lemma 3 is also true for the full-rank Gaussian variational family. The last conclusion implies $\delta_n^{-1}(\hat{\theta}^*_n - \theta_0) \xrightarrow{d} X$ if $\Delta_{n,\theta_0} \xrightarrow{d} X$ for some random variable X. We defer the proof to Appendix F. □

This corollary says that under a full-rank Gaussian variational family, VB is consistent and asymptotically normal in the classical sense. It accurately recovers the asymptotic normal distribution implied by the local asymptotic normality of $M_n(\theta\,;\,x)$.

Before stating the corollary for the factorizable Gaussian variational family, we first present a lemma on the KL minimizer of a Gaussian distribution over the factorizable Gaussian family. We show that the minimizer keeps the mean but has a diagonal covariance matrix that matches the precision. We also show the minimizer has a smaller entropy than the original distribution. This echoes the well-known phenomenon of VB algorithms underestimating the variance.

Lemma 8. The factorizable KL minimizer of a Gaussian distribution keeps the mean and matches the precision:

$$\underset{\mu_0\in\mathbb{R}^d,\,\Sigma_0\in\mathrm{diag}(d\times d)}{\arg\min}\ \mathrm{KL}(\mathcal{N}(\cdot;\mu_0,\Sigma_0)||\mathcal{N}(\cdot;\mu_1,\Sigma_1)) = \mu_1, \Sigma_1^*,$$

where Σ_1^* is diagonal with $\Sigma_{1,ii}^* = ((\Sigma_1^{-1})_{ii})^{-1}$ for $i = 1, 2, \ldots, d$. Hence, the entropy of the factorizable KL minimizer is smaller than or equal to that of the original distribution:

$$\mathbb{H}(\mathcal{N}(\cdot;\mu_0,\Sigma_1^*)) \leq \mathbb{H}(\mathcal{N}(\cdot;\mu_0,\Sigma_1)).$$

Proof sketch of Lemma 8. The first statement is a consequence of a technical calculation of the KL divergence between two normal distributions. We differentiate the KL divergence over μ_0 and the diagonal terms of Σ_0 and obtain the result. The second statement is due to the inequality of the determinant of a positive matrix being always smaller than or equal to the product of its diagonal terms (Amir-Moez and Johnston 1969; Beckenbach and Bellman 2012). In this sense, mean field variational inference underestimates posterior variance. See Appendix G for details. □

The next corollary studies the VB posterior and the VBE under a factorizable Gaussian variational family.

Corollary 9. Posit a factorizable Gaussian variational family,

$$\mathcal{Q}^d = \{q : q(\theta) = \mathcal{N}(m, \Sigma)\}, \tag{17}$$

where Σ positive definite and diagonal. Then

1. $q^*(\theta) \xrightarrow{d} \delta_{\theta_0}$, almost surely under $\mathbb{P}_{\theta_0}$.
2. $||q^*_{\tilde{\theta}}(\cdot) - \mathcal{N}(\cdot\,;\Delta_{n,\theta_0}, V'^{-1}_{\theta_0})||_{\mathrm{TV}} \xrightarrow{P_{\theta_0}} 0$, where V' is diagonal and has the same diagonal entries as V_{θ_0}.
3. $\hat{\theta}_n^* \xrightarrow{\text{a.s.}} \theta_0$.
4. $\delta_n^{-1}(\hat{\theta}_n^* - \theta_0) - \Delta_{n,\theta_0} = o_{P_{\theta_0}}(1)$.

Proof of Corollary 9. This is a direct consequence of Lemma 8, Theorem 5, and Theorem 6. □

This corollary says that under the factorizable Gaussian variational family, VB is consistent and asymptotically normal in the classical sense. The rescaled asymptotic distribution for $\tilde{\theta}$ recovers the mean but underestimates the covariance. This underdispersion is a common phenomenon we see in mean field variational Bayes.

As we mentioned, the VB posterior is underdispersed. One consequence of this property is that its credible sets can suffer from under-coverage. In the literature on VB, there are two main ways to correct this inadequacy. One way is to increase the expressiveness of the variational family $\mathcal{Q}$ to one that accounts for dependencies among latent variables. This approach is taken by much of the recent VB literature, for example, Tran et al. (2015a); Tran, Ranganath, and Blei (2015b); Ranganath, Tran, and Blei (2016b); Ranganath et al. (2016a); Liu and Wang (2016). As long as the expanded variational family $\mathcal{Q}$ contains the mean field family, Theorem 5 and Theorem 6 remain true.

Alternative methods to handling underdispersion center around sensitivity analysis and bootstrap. Giordano, Broderick, and Jordan (2017a) identified the close relationship between Bayesian sensitivity and posterior covariance. They estimated the covariance with the sensitivity of the VB posterior means with respect to perturbations of the data. Chen, Wang, and Erosheva (2017) explored the use of bootstrap in assessing the uncertainty of a variational point estimate. They also studied the underlying bootstrap theory. Giordano et al. (2017b) assessed the clustering stability in Bayesian nonparametric models based on an approximation to the infinitesimal jackknife.

3.4. The LAN Condition of the Variational Log-Likelihood

Our results rest on Assumption 1.3, the LAN expansion of the variational log-likelihood $M_n(\theta\,;\,x)$. For models without local latent variables z, their variational log-likelihood $M_n(\theta\,;\,x)$ is the same as their log-likelihood $\log p(x|\theta)$. The LAN expansion for these models have been widely studied. In particular, iid sampling from a regular parametric model is locally asymptotically normal; it satisfies Assumption 1.3 (Van der Vaart 2000). When models do contain local latent variables, however, as we will see in Section 4, finding the LAN expansion requires model-specific characterization.

For a certain class of models with local latent variables, the LAN expansion for the (complete) log-likelihood $\log p(x, z|\theta)$ concurs with the expansion of the variational log-likelihood $M_n(\theta\,;\,x)$. Below we provide a sufficient condition for such a shared LAN expansion. It is satisfied, for example, by the stochastic block model (Bickel et al. 2013) under mild identifiability conditions.

Proposition 10. The log-likelihood $\log p(x, z|\theta)$ and the variational log likelihood $M_n(\theta\,;\,x)$ will have the same LAN expansion if:

1. The conditioned nuisance posterior is consistent under δ_n-perturbation at some rate ρ_n with $\rho_n \downarrow 0$ and $\delta_n^{-2}\rho_n \to 0$:
   For all bounded, stochastic $h_n = O_{P_{\theta_0}}(1)$, the conditional nuisance posterior converges as
   $$\int_{D^c(\theta,\rho_n)} p(z|x, \theta = \theta_0 + \delta_n h_n)\,dz = o_{P_{\theta_0}}(1),$$
   where $D^c(\theta, \rho_n) = \{z : d_H(z, z_{\text{profile}}) \geq \rho_n\}$ is the Hellinger ball of radius ρ_n around $z_{\text{profile}} = \arg\max_z p(x, z|\theta)$, the maximum profile likelihood estimate of z.
2. ρ_n should also satisfy that the likelihood ratio is dominated:
   $$\sup_{z\in\{z:d_H(z,z_{\text{profile}})<\rho_n\}} \mathbb{E}_{\theta_0}\frac{p(x, z|\theta_0 + \delta_n h_n)}{p(x, z|\theta_0)} = O(1),$$
   where the expectation is taken over x.

Proof sketch of Proposition 10. The first condition roughly says the posterior of the local latent variables z contracts faster than the global latent variables θ. The second condition is a regularity condition. The two conditions together ensure the log marginal likelihood $\log \int p(x, z|\theta)\, dz$ and the complete log-likelihood $\log p(x, z|\theta)$ share the same LAN expansion. This condition shares a similar flavor with the condition (3.1) of the semiparametric Bernstein–von Mises theorem in Bickel et al. (2012). This implication can be proved by a slight adaptation of the proof of Theorem 4.2 in Bickel et al. (2012): We view the collection of local latent variables z as an infinite-dimensional nuisance parameter.

This proposition is due to the following key inequality. For simplicity, we state the version with only discrete local latent variables z:

$$\log p(x, z|\theta) \leq M_n(\theta\, ;\, x) \leq \log \int p(x, z|\theta)\, dz. \tag{18}$$

The continuous version of this inequality can be easily adapted. The lower bound is due to

$$p(x, z|\theta) = \int q(z) \log \frac{p(x, z|\theta)}{q(x)}\, dz \bigg|_{q(z)=\delta_z},$$

and

$$M_n(\theta\, ;\, x) = \sup_{q \in \mathcal{Q}^d} \int q(z) \log \frac{p(x, z|\theta)}{q(x)}\, dz.$$

The upper bound is due to the Jensen's inequality. This inequality ensures that the same LAN expansion for the leftmost and the rightmost terms would imply the same LAN expansion for the middle term, the variational log-likelihood $M_n(\theta\, ;\, x)$. See Appendix H for details. □

In general, we can appeal to Theorem 4 of Le Cam and Yang (2012) to argue for the preservation of the LAN condition, showing that if it holds for the complete log-likelihood then it holds for the variational log-likelihood too. In their terminology, we need to establish the VFE as a "distinguished" statistic.

4. Applications

We proved consistency and asymptotic normality of the variational Bayes (VB) posterior (in total variation (TV) distance) and the variational Bayes estimate (VBE). We mainly relied on the prior mass condition, the local asymptotic normality of the variational log-likelihood $M_n(x\, ;\, \theta)$ and the consistent testability assumption of the data-generating parameter.

We now apply this argument to three types of Bayesian models: Bayesian mixture models (Bishop 2006; Murphy 2012), Bayesian generalized linear mixed models (McCulloch and Neuhaus 2001; Jiang 2007), and Bayesian stochastic block models (Wang and Wong 1987; Snijders and Nowicki 1997; Hofman and Wiggins 2008; Mossel, Neeman, and Sly 2012; Abbe and Sandon 2015). For each model class, we illustrate how to leverage the known asymptotic results for frequentist variational approximations to prove asymptotic results for VB. We assume the prior mass condition for the rest of this section: the prior measure of a parameter θ with Lebesgue density $p(\theta)$ on Θ is continuous and positive on a neighborhood of the true data-generating value θ_0. For simplicity, we posit a mean field family for the local latent variables and a factorizable Gaussian variational family for the global latent variables.

4.1. Bayesian Mixture Models

The Bayesian mixture model is a versatile class of models for density estimation and clustering (Bishop 2006; Murphy 2012).

Consider a Bayesian mixture of K unit-variance univariate Gaussians with means $\mu = \{\mu_1, \ldots, \mu_K\}$. For each observation x_i, $i = 1, \ldots, n$, we first randomly draw a cluster assignment c_i from a categorical distribution over $\{1, \ldots, K\}$; we then draw x_i randomly from a unit-variance Gaussian with mean μ_{c_i}. The model is

$$\begin{aligned} \mu_k &\sim p_\mu, & k &= 1, \ldots, K, \\ c_i &\sim \text{Categorical}(1/K, \ldots, 1/K), & i &= 1, \ldots, n, \\ x_i|c_i, \mu &\sim \mathcal{N}(c_i^\top \mu, 1). & i &= 1, \ldots, n. \end{aligned}$$

For a sample of size n, the joint distribution is

$$p(\mu, c, x) = \prod_{i=1}^{K} p_\mu(\mu_i) \prod_{i=1}^{n} p(c_i) p(x_i|c_i, \mu).$$

Here μ is a K-dimensional global latent vector and $c_{1:n}$ are local latent variables. We are interested inferring the posterior of the μ vector.

We now establish asymptotic properties of VB for Bayesian Gaussian mixture model (GMM).

Corollary 11. Assume the data-generating measure P_{μ_0} has density $\int p(\mu_0, c, x)\, dc$. Let $q^*(\mu)$ and μ^* denote the VB posterior and the VBE. Under regularity conditions (A1–A5) and (B1,2,4) of Westling and McCormick (2015), we have

$$\left\| q^*(\mu) - \mathcal{N}\left(\mu_0 + \frac{Y}{\sqrt{n}}, \frac{1}{n} V_0(\mu_0)\right) \right\|_{\mathrm{TV}} \xrightarrow{P_{\mu_0}} 0,$$

and

$$\sqrt{n}(\mu^* - \mu_0) \xrightarrow{d} Y,$$

where μ_0 is the true value of μ that generates the data. We have

$$\begin{aligned} Y &\sim \mathcal{N}(0, V(\mu_0)), \\ V(\mu_0) &= A(\mu_0)^{-1} B(\mu_0) A(\mu_0)^{-1}, \\ A(\mu) &= \mathbb{E}_{P_{\mu_0}}[D_\mu^2 m(\mu\, ;\, x)], \\ B(\mu) &= \mathbb{E}_{P_{\mu_0}}[D_\mu m(\mu\, ;\, x) D_\mu m(\mu\, ;\, x)^\top], \\ m(\mu\, ;\, x) &= \sup_{q(c) \in \mathcal{Q}^n} \int q(c) \log \frac{p(x, c|\mu)}{q(c)}\, dc. \end{aligned}$$

The diagonal matrix $V_0(\mu_0)$ satisfies $(V_0(\mu_0)^{-1})_{ii} = (A(\mu_0))_{ii}$. The specification of Gaussian mixture model is invariant to permutation among K components; this corollary is true up to permutations among the K components.

Proof sketch for Corollary 11. The consistent testability condition is satisfied by the existence of a consistent estimate due to Theorem 1 of Westling and McCormick (2015). The local asymptotic normality is proved by a Taylor expansion of $m(\mu\, ;\, x)$ at μ_0. This result then follows directly from our Theorem 5 and Theorem 6 in Section 3. The technical conditions inherited from Westling and McCormick (2015) allow us to use their Theorems 1and 2 for properties around variational frequentist estimate (VFE). See Appendix I for proof details. □

4.2. Bayesian Generalized Linear Mixed Models

Bayesian generalized linear mixed model (GLMMs) are a powerful class of models for analyzing grouped data or longitudinal data (McCulloch and Neuhaus 2001; Jiang 2007).

Consider a Poisson mixed model with a simple linear relationship and group-specific random intercepts. Each observation reads (X_{ij}, Y_{ij}), $1 \leq i \leq m$, $1 \leq j \leq n$, where the Y_{ij}'s are nonnegative integers and the X_{ij}'s are unrestricted real numbers. For each group of observations (X_{ij}, Y_{ij}), $1 \leq j \leq n$, we first draw the random effect U_i independently from $N(0, \sigma^2)$. We follow by drawing Y_{ij} from a Poisson distribution with mean $\exp(\beta_0 + \beta_1 X_{ij} + U_i)$. The probability model is

$$\begin{aligned}
\beta_0 &\sim p_{\beta_0}, \\
\beta_1 &\sim p_{\beta_1}, \\
\sigma^2 &\sim p_{\sigma^2}, \\
U_i &\overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2), \\
Y_{ij} | x_{ij}, U_i &\sim \mathrm{Poi}(\exp(\beta_0 + \beta_1 X_{ij} + U_i)).
\end{aligned}$$

The joint distribution is

$$\begin{aligned}
&p(\beta_0, \beta_1, \sigma^2, U_{1:m}, Y_{1:m,1:n} | x_{1:m,1:n}) \\
&\quad = p_{\beta_0}(\beta_0) p_{\beta_1}(\beta_1) p_{\sigma^2}(\sigma^2) \prod_{i=1}^{m} \mathcal{N}(U_i; 0, \sigma^2) \\
&\qquad \times \prod_{i=1}^{m} \prod_{j=1}^{n} \mathrm{Poi}(Y_{ij}; \exp(\beta_0 + \beta_1 X_{ij} + U_i)).
\end{aligned}$$

We establish asymptotic properties of VB in Bayesian Poisson linear mixed models.

Corollary 12. Consider the true data-generating distribution $P_{\beta_0^0, \beta_1^0, (\sigma^2)^0}$ with the global latent variables taking the true values $\{\beta_0^0, \beta_1^0, (\sigma^2)^0\}$. Let $q^*_{\beta_0}(\beta_0)$, $q^*_{\beta_1}(\beta_1)$, $q^*_{\sigma^2}(\sigma^2)$ denote the VB posterior of $\beta_0, \beta_1, \sigma^2$. Similarly, let $\beta_0^*, \beta_1^*, (\sigma^2)^*$ be the VBEs accordingly. Consider $m = O(n^2)$. Under regularity conditions (A1–A5) of Hall et al. (2011b), we have

$$\begin{aligned}
&\Bigg\| q^*_{\beta_0}(\beta_0) q^*_{\beta_1}(\beta_1) q^*_{\sigma^2}(\sigma^2) - \mathcal{N}\Bigg(\left(\beta_0^0, \beta_1^0, (\sigma^2)^0\right) \\
&\quad + \left(\frac{Z_1}{\sqrt{n}}, \frac{Z_2}{\sqrt{mn}}, \frac{Z_3}{\sqrt{n}} \right), \mathrm{diag}(V_1, V_2, V_3) \Bigg) \Bigg\|_{\mathrm{TV}} \overset{P_{\beta_0^0, \beta_1^0, (\sigma^2)^0}}{\rightarrow} 0,
\end{aligned}$$

where

$$\begin{aligned}
&Z_1 \sim \mathcal{N}(0, (\sigma^2)^0), Z_2 \sim \mathcal{N}(0, \tau^2), Z_3 \sim \mathcal{N}(0, 2\{(\sigma^2)^0\}^2), \\
&V_1 = \exp\left(-\beta_0 + \frac{1}{2}(\sigma^2)^0 \right) / \phi\left(\beta_1^0\right), \\
&V_2 = \exp\left(-\beta_0^0 + \frac{1}{2}\sigma^2 \right) / \phi''\left(\beta_1^0\right), \\
&V_3 = 2\{(\sigma^2)^0\}^2, \\
&\tau^2 = \frac{\exp\{-(\sigma^2)^0/2 - \beta_0^0\} \phi(\beta_1^0)}{\phi''(\beta_1^0)\phi(\beta_1^0) - \phi'(\beta_1^0)^2}.
\end{aligned}$$

Here $\phi(\cdot)$ is the moment-generating function of X.

Also,

$$\begin{aligned}
&\left(\sqrt{m}\left(\beta_0^* - \beta_0^0\right), \sqrt{mn}\left(\beta_1^* - \beta_1^0\right),\right. \\
&\quad \left.\sqrt{m}((\sigma^2)^* - (\sigma^2)^0)\right) \xrightarrow{d} (Z_1, Z_2, Z_3).
\end{aligned}$$

Proof sketch for Corollary 12. The consistent testability assumption is satisfied by the existence of consistent estimates of the global latent variables shown in Theorem 3.1 of Hall et al. (2011b). The local asymptotic normality is proved by a Taylor expansion of the variational log-likelihood based on estimates of the variational parameters based on eqs. (5.18) and (5.22) of Hall et al. (2011b). The technical conditions inherited from Hall et al. (2011b) allow us to leverage their Theorem 3.1 for properties of the VFE. The result then follows directly from Theorem 5 and Theorem 6 in Section 3. See Appendix K for proof details. □

4.3. Bayesian Stochastic Block Models

Stochastic block models are an important methodology for community detection in network data (Wang and Wong 1987; Snijders and Nowicki 1997; Mossel, Neeman, and Sly 2012; Abbe and Sandon 2015).

Consider n vertices in a graph. We observe pairwise linkage between nodes $A_{ij} \in \{0, 1\}$, $1 \leq i, j \leq n$. In a stochastic block model, this adjacency matrix is driven by the following process: first assign each node i to one of the K latent classes by a categorical distribution with parameter π. Denote the class membership as $Z_i \in \{1, \ldots, K\}$. Then draw $A_{ij} \sim \mathrm{Bernoulli}(H_{Z_i, Z_j})$. The parameter H is a symmetric matrix in $[0, 1]^{K \times K}$ that specifies the edge probabilities between two latent classes; the parameter π are the proportions of the latent classes. The Bayesian stochastic block model is

$$\begin{aligned}
\pi &\sim p(\pi), \\
H &\sim p(H), \\
Z_i | \pi &\overset{\text{iid}}{\sim} \mathrm{Categorical}(\pi), \\
A_{ij} | Z_i, Z_j, H &\overset{\text{iid}}{\sim} \mathrm{Bernoulli}(H_{Z_i Z_j}).
\end{aligned}$$

The dependence in stochastic block model is more complicated than the Bayesian GMM or the Bayesian GLMM.

Before establishing the result, we reparameterize (π, H) by $\theta = (\omega, \nu)$, where $\omega \in \mathbb{R}^{K-1}$ is the log odds ratio of belonging to classes $1, \ldots, K-1$, and $\nu \in \mathbb{R}^{K \times K}$ is the log odds ratio of an edge existing between all pairs of the K classes. The reparameterization is

$$\begin{aligned}
\omega(a) &= \log \frac{\pi(a)}{1 - \sum_{b=1}^{K-1} \pi(b)}, \qquad a = 1, \ldots, K-1, \\
\nu(a, b) &= \log \frac{H(a, b)}{1 - H(a, b)}, \qquad a, b = 1, \ldots, K.
\end{aligned}$$

The joint distribution is

$$\begin{aligned}
p(\theta, Z, A) &= \prod_{a=1}^{K-1} \left[e^{\omega(a) n_a} \left(1 + \sum_{a=1}^{K-1} e^{\omega(a)} \right)^{-n} \right] \\
&\quad \times \prod_{a=1}^{K} \prod_{b=1}^{K} [e^{\nu(a,b) O_{ab}} (1 + e^{\nu(a,b)})^{n_{ab}}]^{1/2},
\end{aligned}$$

where

$$\begin{aligned} n_a(Z) &= \sum_{i=1}^{n} 1\{Z_i = a\}, \\ n_{ab}(Z) &= \sum_{i=1}^{n} \sum_{j \neq i}^{n} 1\{Z_i = a, Z_j = b\}, \\ O_{ab}(A, Z) &= \sum_{i=1}^{n} \sum_{j \neq i}^{n} 1\{Z_i = a, Z_j = b\} A_{ij}. \end{aligned}$$

We now establish the asymptotic properties of VB for stochastic block models.

Corollary 13. Consider ν_0, ω_0 as true data-generating parameters. Let $q_\nu^*(\nu)$, $q_\omega^*(\omega)$ denote the VB posterior of ν and ω. Similarly, let ν^*, ω^* be the VBE. Then

$$\Bigg\| q_\nu^*(\nu) q_\omega^*(\omega) - \mathcal{N}\left((\nu, \omega); (\nu_0, \omega_0) + \left(\frac{\Sigma_1^{-1} Y_1}{\sqrt{n\lambda_0}}, \frac{\Sigma_2^{-1} Y_2}{\sqrt{n}} \right), V_n(\nu_0, \omega_0) \right) \Bigg\|_{\mathrm{TV}} \overset{P_{\nu_0,\omega_0}}{\to} 0,$$

where $\lambda_0 = \mathbb{E}_{P_{\nu_0,\omega_0}}(\text{degree of each node})$, $(\log n)^{-1}\lambda_0 \to \infty$. Y_1 and Y_2 are two zero mean random vectors with covariance matrices Σ_1 and Σ_2, where Σ_1, Σ_2 are known functions of ν_0, ω_0. The diagonal matrix $V(\nu_0, \omega_0)$ satisfies $V^{-1}(\nu_0, \omega_0)_{ii} = \mathrm{diag}(\Sigma_1, \Sigma_2)_{ii}$. Also,

$$\left(\sqrt{n\lambda_0}(\nu^* - \nu_0), \sqrt{n}(\omega^* - \omega_0) \right) \overset{d}{\to} \left(\Sigma_1^{-1} Y_1, \Sigma_2^{-1} Y_2 \right),$$

The specification of classes in stochastic block model (SBM) is permutation invariant. So the convergence above is true up to permutation with the K classes. We follow Bickel et al. (2013) to consider the quotient space of (ν, ω) over permutations.

Proof sketch of Corollary 13. The consistent testability assumption is satisfied by the existence of consistent estimates by Lemma 1 of Bickel et al. (2013). The local asymptotic normality,

$$\begin{aligned} &\sup_{q(z) \in \mathcal{Q}^K} \int q(z) \log \frac{p(A, z | \nu_0 + \frac{t}{\sqrt{n^2 \rho_n}}, \omega_0 + \frac{s}{\sqrt{n}})}{q(z)} \, dz \\ &= \sup_{q(z) \in \mathcal{Q}^K} \int q(z) \log \frac{p(A, z | \nu_0, \omega_0)}{q(z)} \, dz \\ &\quad + s^\top Y_1 + t^\top Y_2 - \frac{1}{2} s^\top \Sigma_1 s - \frac{1}{2} t^\top \Sigma_2 t + o_P(1), \quad (19) \end{aligned}$$

for $(\nu_0, \omega_0) \in \mathcal{T}$ for compact $\mathcal{T}$ with $\rho_n = \frac{1}{n}\mathbb{E}(\text{degree of each node})$, is established by Lemma 2, Lemma 3, and Theorem 3 of Bickel et al. (2013). The result then follows directly from our Theorem 5 and Theorem 6 in Section 3. See Appendix K for proof details. □

5. Simulation Studies

We illustrate the implication of Theorem 5 and Theorem 6 by simulation studies on Bayesian GLMM (McCullagh 1984). We also study the VB posteriors of latent Dirichlet allocation (LDA) (Blei, Ng, and Jordan 2003). This is a model that shares similar structural properties with SBM but has no consistency results established for its VFE.

We use two automated inference algorithms offered in Stan, a probabilistic programming system (Carpenter et al. 2015): VB through automatic differentiation variational inference (ADVI) (Kucukelbir et al. 2017) and Hamiltonian Monte Carlo (HMC) simulation through No-U-Turn sampler (NUTS) (Hoffman and Gelman 2014). We note that optimization algorithms used for VB in practice only find local optima.

In both cases, we observe the VB posteriors get closer to the truth as the sample size increases; when the sample size is large enough, they coincide with the truth. They are underdispersed, however, compared with HMC methods.

5.1. *Bayesian Generalized Linear Mixed Models*

We consider the Poisson linear mixed model studied in Section 4. Fix the group size as $n = 10$. We simulate datasets of size $N = (50, 100, 200, 500, 1000, 2000, 5000, 10{,}000, 20{,}000)$. As the size of the dataset grows, the number of groups also grows; so does the number of local latent variables U_i, $1 \leq i \leq m$. We generate a four-dimensional covariate vector for each X_{ij}, $1 \leq i \leq m$, $1 \leq j \leq n$, where the first dimension follows iid $\mathcal{N}(0, 1)$, the second dimension follows iid $\mathcal{N}(0, 25)$, the third dimension follows iid $\mathrm{Bernoulli}(0.4)$, and the fourth dimension follows iid $\mathrm{Bernoulli}(0.8)$. We wish to study the behaviors of coefficient efficients for underdispersed/overdispersed continuous covariates and balanced/imbalanced binary covariates. We set the true parameters as $\beta_0 = 5$, $\beta_1 = (0.2, -0.2, 2, -2)$, and $\sigma^2 = 2$.

Figure 1 shows the boxplots of VB posteriors for β_0, β_1, and σ^2. All VB posteriors converge to their corresponding true values as the size of the dataset increases. The box plots present rather few outliers; the lower fence, the box, and the upper fence are about the same size. This suggests normal VB posteriors. This echoes the consistency and asymptotic normality concluded from Theorem 5. The VB posteriors are underdispersed, compared to the posteriors via HMC. This also echoes our conclusion of underdispersion in Theorem 5 and Lemma 8.

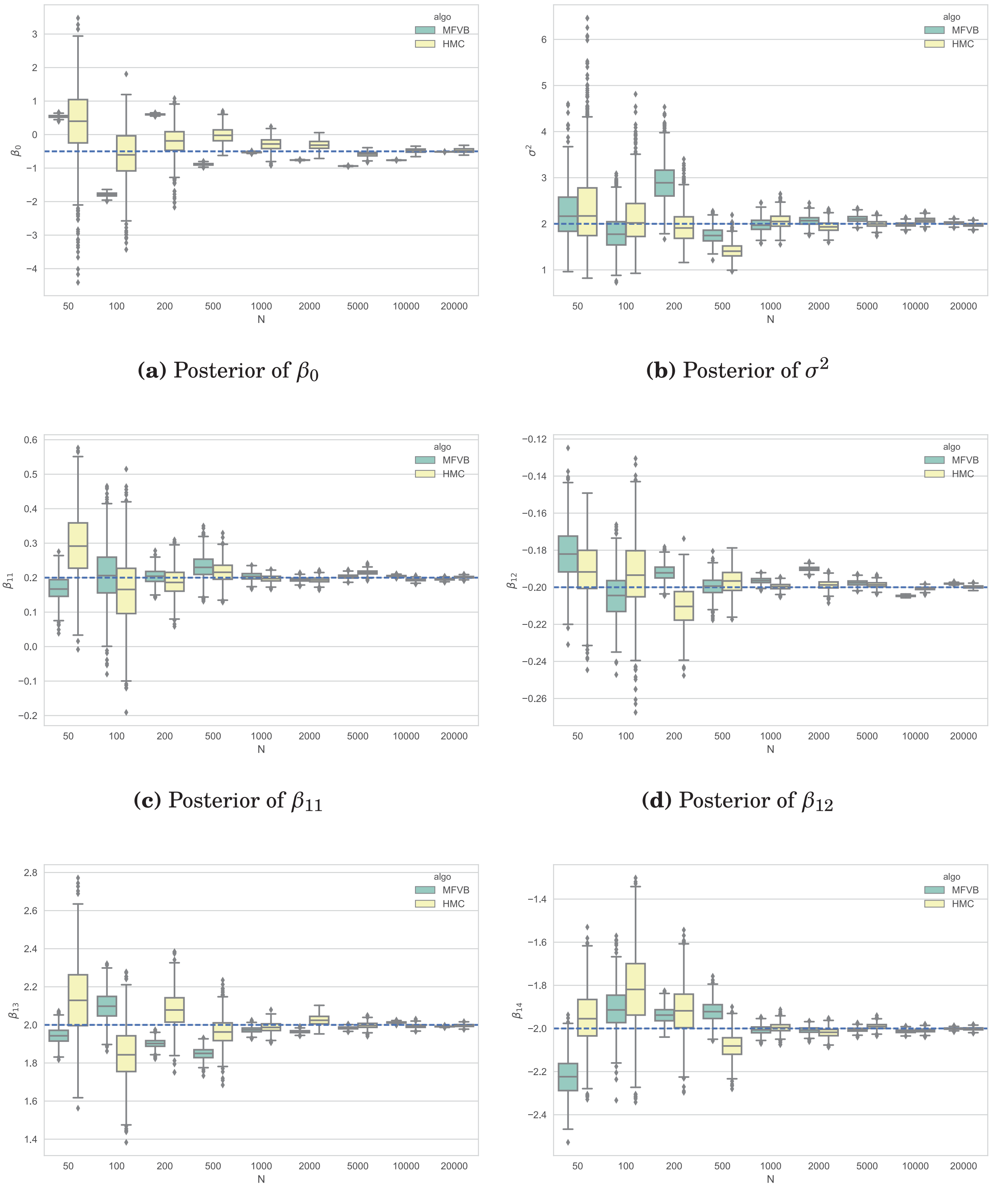
Regarding the convergence rate, VB posteriors of all dimensions of β_1 quickly converge to their true value; the VB posteriors center around their true values as long as $N \geq 1000$. The convergence of VB posteriors of slopes for continuous variables (β_{11}, β_{12}) are generally faster than those for binary ones (β_{13}, β_{14}). The VB posterior of σ^2 shares a similarly fast convergence rate. The VB posterior of the intercept β_0, however, struggles; it is away from the true value until the dataset size hits $N = 20,000$. This aligns with the convergence rate inferred in Corollary 12, $\sqrt{mn}$ for β_1 and $\sqrt{m}$ for β_0 and σ^2.

Computation wise, VB takes orders of magnitude less time than HMC. The performance of VB posteriors is comparable with that from HMC when the sample size is sufficiently large; in this case, we need $N = 20,000$.

5.2. *Latent Dirichlet Allocation*

Latent Dirichlet allocation (LDA) is a generative statistical model commonly adopted to describe word distributions in documents by latent topics.

Given M documents, each with N_m, $m = 1, \ldots, M$ words, composing a vocabulary of V words, we assume K latent topics. Consider two sets of latent variables: topic distributions for



(a) Posterior of β_0

(b) Posterior of σ^2

(c) Posterior of β_{11}

(d) Posterior of β_{12}

(e) Posterior of β_{13}

(f) Posterior of β_{14}

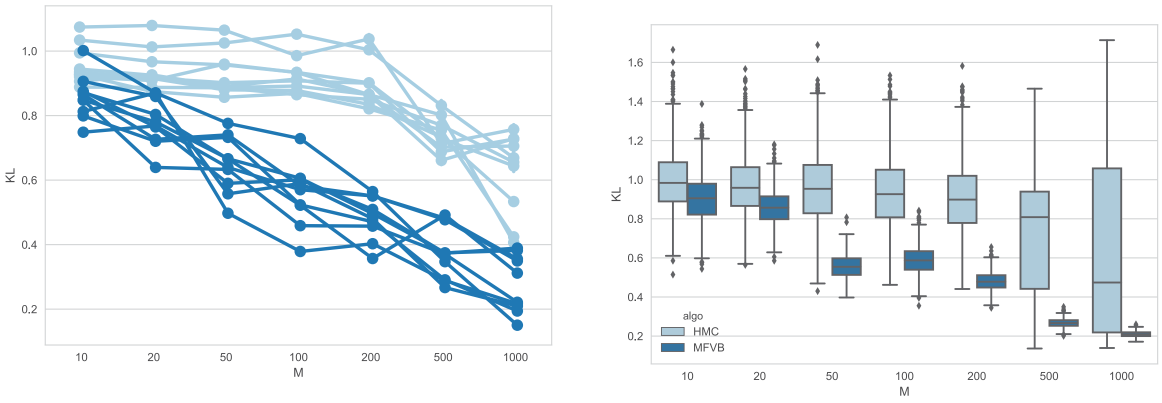
Figure 1. VB posteriors and HMC posteriors of Poisson generalized linear mixed model versus size of datasets. VB posteriors are consistent and asymptotically normal but underdispersed than HMS posteriors. β_0 and σ^2 converge to the truth slower than β_1 does. They echo our conclusions in Theorems 5 and Corollary 12.

document m, $(\theta_m)_{K\times 1}$, $m = 1, \dots, M$ and word distributions for topic k, $(\phi_k)_{V\times 1}$, $k = 1, \dots, K$. The generative process is

$$\begin{aligned} \theta_m &\sim p_\theta, & m &= 1, \dots, M, \\ \phi_k &\sim p_\phi, & k &= 1, \dots, K, \\ z_{m,j} &\sim \text{Mult}(\theta_m), & j &= 1, \dots, N_m, m = 1, \dots, M, \\ w_{m,j} &\sim \text{Mult}(\phi_{z_{m,j}}), & j &= 1, \dots, N_m, m = 1, \dots, M. \end{aligned}$$

The first two rows are assigning priors assigned to the latent variables. $w_{m,j}$ denotes word j of document m and $z_{m,j}$ denotes its assigned topic.

We simulate a dataset with $V = 100$ sized vocabulary and $K = 10$ latent topics in $M = (10, 20, 50, 100, 200, 500, 100)$ documents. Each document has N_m words where $N_m \overset{\text{iid}}{\sim} \text{Poi}(100)$. As



(a) Posterior mean KL divergence of the $K = 10$ topics

(b) Boxplots of posterior KL divergence of Topic 2

Figure 2. Mean of Kullback–Leibler (KL) divergence between the true topics and the fitted VB and HMC posterior topics versus size of datasets. (a) VB posteriors (dark blue) converge to the truth; they are very close to the truth as we hit $M = 1000$ documents. (b) VB posteriors are consistent but underdispersed compared to HMC posteriors (light blue). These align with our conclusions in Theorem 5.

the number of documents M grows, the number of document-specific topic vectors θ_m grows while the number of topic-specific word vectors ϕ_k stays the same. In this sense, we consider $\theta_m, m = 1, \dots, M$ as local latent variables and $\phi_k, k = 1, \dots, K$ as global latent variables. We are interested in the VB posteriors of global latent variables $\phi_k, k = 1, \dots, K$ here. We generate the datasets with true values of θ and ϕ, where they are random draws from $\theta_m \overset{\text{iid}}{\sim} \text{Dir}((1/K)_{K\times 1})$ and $\phi_k \overset{\text{iid}}{\sim} \text{Dir}((1/V)_{V\times 1})$.

Figure 2 presents the KL divergence between the $K = 10$ topic-specific word distributions induced by the true ϕ_k's and the fitted ϕ_k's by VB and HMC. This KL divergence equals to $\text{KL}(\text{Mult}(\phi_k^0)||\text{Mult}(\hat{\phi}_k)) = \sum_{i=1}^V \phi_{ki}^0(\log \phi_{ki}^0 - \log \hat{\phi}_{ki})$, where ϕ_{ki}^0 is the ith entry of the true kth topic and $\hat{\phi}_{ki}$ is the ith entry of the fitted kth topic.

Figure 2(a) shows that VB posterior (dark blue) mean KL divergences of all $K = 10$ topics get closer to 0 as the number of documents M increase, faster than HMC (light blue). We become very close to the truth as the number of documents M hits 1000. Figure 2(b) (we only show boxplots for Topic 2 here. The boxplots of other topics look very similar) shows that the boxplots of VB posterior mean KL divergences get closer to 0 as M increases. They are underdispersed compared to HMC posteriors. These align with our understanding of how VB posterior behaves in Theorem 5.

Computation wise, again VB is orders of magnitude faster than HMC. In particular, optimization in VB in our simulation studies converges within 10,000 steps.

6. Discussion

Variational Bayes (VB) methods are a fast alternative to Markov chain Monte Carlo (MCMC) for posterior inference in Bayesian modeling. However, few theoretical guarantees have been established. This work proves consistency and asymptotic normality for variational Bayes (VB) posteriors. The convergence is in the sense of total variation (TV) distance converging to zero in probability. In addition, we establish consistency and asymptotic normality of variational Bayes estimate (VBE). The result is frequentist in the sense that we assume a data-generating distribution driven by some fixed nonrandom true value for global latent variables.

These results rest on ideal variational Bayes and its connection to frequentist variational approximations. Thus this work bridges the gap in asymptotic theory between the frequentist variational approximation, in particular the variational frequentist estimate (VFE), and variational Bayes. It also assures us that variational Bayes as a popular approximate inference algorithm bears some theoretical soundness.

We present our results in the classical VB framework but the results and proof techniques are more generally applicable. Our results can be easily generalized to more recent developments of VB beyond Kullback–Leibler (KL) divergence, α-divergence, or χ-divergence, for example (Li and Turner 2016; Dieng et al. 2017). They are also applicable to more expressive variational families, as long as they contain the mean field family. We could also allow for model misspecification, as long as the variational log-likelihood $M_n(\theta\ ;\ x)$ under the misspecified model still enjoys local asymptotic normality.

There are several interesting avenues for future work. The variational Bernstein–von Mises theorem developed in this work applies to parametric and semiparametric models. One direction is to study the VB posteriors in nonparametric settings. A second direction is to characterize the finite-sample properties of VB posteriors. Finally, we characterized the asymptotics of an optimization problem, assuming that we obtain the global optimum. Though our simulations corroborated the theory, VB optimization typically finds a local optimum. Theoretically characterizing these local optima requires further study of the optimization loss surface.

Supplementary Materials

The online supplementary materials contain the appendices for the article.

Acknowledgments

The authors thank the associate editor and two anonymous reviewers for their constructive comments. The authors thank Adji Dieng, Christian

Naesseth, and Dustin Tran for their valuable feedback on their manuscript. The authors also thank Richard Nickl for pointing them to a key reference.

Funding

This work is supported by ONR N00014-11-1-0651, DARPA PPAML FA8750-14-2-0009, the Alfred P. Sloan Foundation, and the John Simon Guggenheim Foundation.

ORCID

Yixin Wang http://orcid.org/0000-0002-6617-4842

References

Abbe, E., and Sandon, C. (2015), "Community Detection in General Stochastic Block Models: Fundamental Limits and Efficient Recovery Algorithms," in *2015 IEEE 56th Annual Symposium on Foundations of Computer Science* (FOCS), Piscataway, NJ: IEEE, pp. 670–688. [1155,1156]

Alquier, P., and Ridgway, J. (2017), "Concentration of Tempered Posteriors and of Their Variational Approximations," arXiv:1706.09293. [1149]

Alquier, P., Ridgway, J., and Chopin, N. (2016), "On the Properties of Variational Approximations of Gibbs Posteriors," *Journal of Machine Learning Research*, 17, 1–41. [1149]

Amir-Moez, A., and Johnston, G. (1969), "On the Product of Diagonal Elements of a Positive Matrix," *Mathematics Magazine*, 42, 24–26. [1154]

Beckenbach, E. F., and Bellman, R. (2012), *Inequalities* (Vol. 30), New York: Springer Science & Business Media. [1154]

Bernstein, S. N. (1917), *Theory of Probability*, Moscow, Leningrad. [1150]

Bickel, P., Choi, D., Chang, X., and Zhang, H. (2013), "Asymptotic Normality of Maximum Likelihood and Its Variational Approximation for Stochastic Blockmodels," *The Annals of Statistics*, 41, 1922–1943. [1148,1150,1154,1157]

Bickel, P., and Kleijn, B. (2012), "The Semiparametric Bernstein–von Mises Theorem," *The Annals of Statistics*, 40, 206–237. [1150,1155]

Bickel, P. J., and Yahav, J. A. (1967), "Asymptotically Pointwise Optimal Procedures in Sequential Analysis," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 401–413). [1150]

Bishop, C. M. (2006), "Pattern Recognition," in *Machine Learning*, New York: Springer-Verlag, p. 128. [1153,1155]

Blei, D., Kucukelbir, A., and McAuliffe, J. (2016), "Variational Inference: A Review for Statisticians," *Journal of American Statistical Association*, 112, 859–877. [1147,1148]

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003), "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, 3, 993–1022. [1147,1157]

Bontemps, D. et al., (2011), "Bernstein–Von Mises Theorems for Gaussian Regression With Increasing Number of Regressors," *The Annals of Statistics*, 39, 2557–2584. [1150]

Boucheron, S., Gassiat, E., et al., (2009), "A Bernstein-Von Mises Theorem for Discrete Probability Distributions," *Electronic Journal of Statistics*, 3, 114–148. [1150]

Braides, A. (2006), "A Handbook of Γ-Convergence," in *Handbook of Differential Equations: Stationary Partial Differential Equations* (Vol. 3), eds. M. Chipot and P. Quittner, The Netherlands, Elsevier, pp. 101–213. [1152]

Breslow, N. E., and Clayton, D. G. (1993), "Approximate Inference in Generalized Linear Mixed Models," *Journal of the American Statistical Association*, 88, 9–25. [1147]

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2015), "Stan: A Probabilistic Programming Language," *Journal of Statistical Software*, 76, 1–32. [1157]

Castillo, I. (2012a), "Semiparametric Bernstein–Von Mises Theorem and Bias, Illustrated With Gaussian Process Priors," *Sankhya A*, 74, 194–221. [1150]

——— (2012b), "A Semiparametric Bernstein–Von Mises Theorem for Gaussian Process Priors," *Probability Theory and Related Fields*, 152, 53–99. [1150]

——— (2014), "On Bayesian Supremum Norm Contraction Rates," *The Annals of Statistics*, 42, 2058–2091. [1150]

Castillo, I., and Nickl, R. (2012), "Nonparametric Bernstein–Von Mises Theorems," arXiv:1208.3862. [1150]

——— (2013), "Nonparametric Bernstein–Von Mises Theorems in Gaussian White Noise," *The Annals of Statistics*, 41, 1999–2028. [1150]

——— (2014), "On the Bernstein–von Mises Phenomenon for Nonparametric Bayes Procedures," *The Annals of Statistics*, 42, 1941–1969. [1150]

Castillo, I., and Rousseau, J. (2015), "A Bernstein–Von Mises Theorem for Smooth Functionals in Semiparametric Models," *The Annals of Statistics*, 43, 2353–2383. [1150]

Celisse, A., Daudin, J.-J., and Pierre, L. (2012), "Consistency of Maximum-Likelihood and Variational Estimators in the Stochastic Block Model," *Electronic Journal of Statistics*, 6, 1847–1899. [1149]

Chen, Y.-C., Wang, Y. S., and Erosheva, E. A. (2017), "On the Use of Bootstrap With Variational Inference: Theory, Interpretation, and a Two-Sample Test Example," arXiv:1711.11057. [1154]

Cox, D. D. (1993), "An Analysis of Bayesian Inference for Nonparametric Regression," *The Annals of Statistics*, 21, 903–923. [1150]

Dal Maso, G. (2012), *An Introduction to Γ-Convergence* (Vol. 8), New York: Springer Science & Business Media. [1152]

De Blasi, P., and Hjort, N. L. (2009), "The Bernstein–Von Mises Theorem in Semiparametric Competing Risks Models," *Journal of Statistical Planning and Inference*, 139, 2316–2328. [1150]

Dempster, A., Laird, N., and Rubin, D. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, Series B, 39, 1–38. [1148]

Diaconis, P., and Freedman, D. (1986), "On the Consistency of Bayes Estimates," *The Annals of Statistics*, 14, 1–26. [1150]

——— (1997), "Consistency of Bayes Estimates for Nonparametric Regression: A Review," in *Festschrift for Lucien Le Cam*, eds. D. Pollard, E. Torgersen, and G. L. Yang, New York: Springer, pp. 157–165. [1150]

——— (1998), "Consistency of Bayes Estimates for Nonparametric Regression: Normal Theory," *Bernoulli*, 4, 411–444. [1150]

Dieng, A. B., Tran, D., Ranganath, R., Paisley, J., and Blei, D. M. (2017), "Variational Inference via χ-Upper Bound Minimization," in *Advances in Neural Information Processing Systems*, pp. 2729–2738. [1159]

Freedman, D. (1999), "Wald Lecture: On the Bernstein-Von Mises Theorem With Infinite-Dimensional Parameters," *The Annals of Statistics*, 27, 1119–1141. [1150]

Gelfand, A. E., and Smith, A. F. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409. [1147]

Ghorbani, B., Javadi, H., and Montanari, A. (2018), "An Instability in Variational Inference for Topic Models," arXiv:1802.00568. [1149]

Ghosal, S., and van der Vaart, A. (2017), *Fundamentals of Nonparametric Bayesian Inference* (Vol. 44), Cambridge, UK: Cambridge University Press. [1150]

Ghosh, J., and Ramamoorthi, R. (2003), *Bayesian Nonparametrics* (*Springer Series in Statistics*), New York: Springer. [1150]

Giordano, R., Broderick, T., and Jordan, M. I. (2017a), "Covariances, Robustness, and Variational Bayes," arXiv:1709.02536. [1154]

Giordano, R., Liu, R., Varoquaux, N., Jordan, M. I., and Broderick, T. (2017b), "Measuring Cluster Stability for Bayesian Nonparametrics Using the Linear Bootstrap," arXiv:1712.01435. [1154]

Hall, P., Ormerod, J. T., and Wand, M. (2011a), "Theory of Gaussian Variational Approximation for a Poisson Mixed Model," *Statistica Sinica*, 21, 369–389. [1149]

Hall, P., Pham, T., Wand, M. P., and Wang, S. S. (2011b), "Asymptotic Normality and Valid Inference for Gaussian Variational Approximation," *The Annals of Statistics*, 39, 2502–2532. [1148,1150,1156]

Hastings, W. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109. [1147]

Hoffman, M., Blei, D., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. [1147,1152]

Hoffman, M. D., and Gelman, A. (2014), "The No-U-Turn Sampler," *JMLR*, 15, 1593–1623. [1157]

Hofman, J., and Wiggins, C. (2008), "Bayesian Approach to Network Modularity," *Physical Review Letters*, 100, 258701-1–258701-4. [1147,1155]

James, L. F. (2008), "Large Sample Asymptotics for the Two-Parameter Poisson–Dirichlet Process," in *Pushing the Limits of Contemporary*

Statistics: Contributions in Honor of Jayanta K. Ghosh, eds. B. S. Clarke and S. Ghosal, Beachwood, OH: Institute of Mathematical Statistics, pp. 187–199. [1150]

Jiang, J. (2007), *Linear and Generalized Linear Mixed Models and Their Applications*, New York: Springer Science & Business Media. [1155,1156]

Johnstone, I. M. (2010), "High Dimensional Bernstein-Von Mises: Simple Examples," *Institute of Mathematical Statistics Collections*, 6, 87. [1150]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [1147,1148]

Kim, Y. (2006), "The Bernstein–Von Mises Theorem for the Proportional Hazard Model," *The Annals of Statistics*, 34, 1678–1700. [1150]

——— (2009), "A Bernstein-Von Mises Theorem for Doubly Censored Data," *Statistica Sinica*, 19, 581–595. [1150]

Kim, Y., and Lee, J. (2004), "A Bernstein-Von Mises Theorem in the Nonparametric Right-Censoring Model," *Annals of Statistics*, 32, 1492–1512. [1150]

Kleijn, B., and Van der Vaart, A. (2012), "The Bernstein-von-Mises Theorem Under Misspecification," *Electronic Journal of Statistics*, 6, 354–381. [1150,1151,1153]

Knapik, B. T., van der Vaart, A. W., van Zanten, J. H. et al. (2011), "Bayesian Inverse Problems With Gaussian Priors," *The Annals of Statistics*, 39, 2626–2657. [1150]

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic Differentiation Variational Inference," *The Journal of Machine Learning Research*, 18, 430–474. [1157]

Laplace, P. (1809), "Memoire Sur Les Integrales Definies et leur Application aux Probabilites, et Specialement a la Recherche du Milieu qu'il Faut Choisir Entre les Resultats des Observations," *Memoires Presentes a l'Academie Des Sciences, Paris.* [1150]

Leahu, H. (2011), "On the Bernstein-Von Mises Phenomenon in the Gaussian White Noise Model," *Electronic Journal of Statistics*, 5, 373–404. [1150]

Le Cam, L. (1953), "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes' Estimates," *University of California Publications in Statistics*, 1, 277–330. [1150]

Le Cam, L., and Yang, G. L. (2012), *Asymptotics in Statistics: Some Basic Concepts*, New York: Springer Science & Business Media. [1150,1156]

Lehmann, E. L., and Casella, G. (2006), *Theory of Point Estimation*, New York: Springer Science & Business Media. [1150]

Li, Y., and Turner, R. E. (2016), "Rényi Divergence Variational Inference," in *Advances in Neural Information Processing Systems*, pp. 1073–1081. [1159]

Liu, Q., and Wang, D. (2016), "Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm," in *Advances In Neural Information Processing Systems*, pp. 2378–2386. [1154]

Lu, Y. (2017), "On the Bernstein-Von Mises Theorem for High Dimensional Nonlinear Bayesian Inverse Problems," arXiv:1706.00289. [1150]

Lu, Y., Stuart, A. M., and Weber, H. (2017), "Gaussian Approximations for Probability Measures on $\mathbf{R}^d$," *SIAM/ASA Journal on Uncertainty Quantification*, 5, 1136–1165. [1150,1151]

McCullagh, P. (1984), "Generalized Linear Models," *European Journal of Operational Research*, 16, 285–292. [1157]

McCulloch, C. E., and Neuhaus, J. M. (2001), *Generalized Linear Mixed Models*, New York: Wiley Online Library. [1155,1156]

Mossel, E., Neeman, J., and Sly, A. (2012), "Stochastic Block Models and Reconstruction," arXiv:1202.1499. [1155,1156]

Murphy, K. P. (2012), *Machine Learning: A Probabilistic Perspective*, Cambridge, MA: MIT Press. [1155]

Murphy, S. A., and Van der Vaart, A. W. (2000), "On Profile Likelihood," *Journal of the American Statistical Association*, 95, 449–465. [1150]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [1148]

Ormerod, J. T., You, C., and Muller, S. (2014), "A Variational Bayes Approach to Variable Selection," *Electronic Journal of Statistics*, 11, 3549–3594. [1149]

Panov, M., and Spokoiny, V. (2015), "Finite Sample Bernstein–Von Mises Theorem for Semiparametric Problems," *Bayesian Analysis*, 10, 665–710. [1150]

——— (2014), "Critical Dimension in the Semiparametric Bernsteinvon Mises Theorem," *Proceedings of the Steklov Institute of Mathematics*, 287, 232–255. [1150]

Pati, D., Bhattacharya, A., and Yang, Y. (2017), "On Statistical Optimality of Variational Bayes," arXiv:1712.08983. [1149]

Ranganath, R., Tran, D., Altosaar, J., and Blei, D. (2016a), "Operator Variational Inference," in *Advances in Neural Information Processing Systems*, pp. 496–504. [1154]

Ranganath, R., Tran, D., and Blei, D. (2016b), "Hierarchical Variational Models," in *Proceedings of The 33rd International Conference on Machine Learning*, pp. 324–333. [1154]

Ray, K. (2017), "Adaptive Bernstein-Von Mises Theorems in Gaussian White Noise," *The Annals of Statistics*, 45, 2511–2536. [1150]

Rivoirard, V., and Rousseau, J. (2012), "Bernstein–Von Mises Theorem for Linear Functionals of the Density," *The Annals of Statistics*, 40, 1489–1523. [1150]

Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods* (*Springer Texts in Statistics*), New York: Springer-Verlag. [1147]

Roberts, S. J., Husmeier, D., Rezek, I., and Penny, W. (1998), "Bayesian Approaches to Gaussian Mixture Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 1133–1142. [1147]

Sheth, R., and Khardon, R. (2017), "Excess Risk Bounds for the Bayes Risk Using Variational Inference in Latent Gaussian Models," in *Advances in Neural Information Processing Systems*, pp. 5157–5167. [1149]

Snijders, T. A., and Nowicki, K. (1997), "Estimation and Prediction for Stochastic Blockmodels for Graphs With Latent Block Structure," *Journal of Classification*, 14, 75–100. [1155,1156]

Spokoiny, V. (2013), "Bernstein-Von Mises Theorem for Growing Parameter Dimension," arXiv:1302.3430. [1150]

Tran, D., Blei, D., and Airoldi, E. M. (2015a), "Copula Variational Inference," in *Advances in Neural Information Processing Systems*, pp. 3564–3572. [1154]

Tran, D., Ranganath, R., and Blei, D. M. (2015b), "The Variational Gaussian Process," arXiv:1511.06499. [1154]

Van der Vaart, A. W. (2000), *Asymptotic Statistics* (Vol. 3), Cambridge, UK: Cambridge University Press. [1150,1154]

Von Mises, R. (1931), *Wahrscheinlichkeitsrechnung und ihre Anwendungen in der Statistik und der Theoretischen Physik*, Leipzig und Wien. [1150]

Wainwright, M. J., and Jordan, M. I. (2008), "Graphical Models, Exponential Families, and Variational Inference," *Foundations and Trends® in Machine Learning*, 1, 1–305. [1147,1148]

Wang, B., and Titterington, D. (2004), "Convergence and Asymptotic Normality of Variational Bayesian Approximations for Exponential Family Models With Missing Values," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, AUAI Press, pp. 577–584. [1149]

——— (2005), "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, eds. R. G. Corwell and Z. Ghahramani, pp. 373–380. [1149]

Wang, B., and Titterington, D. (2006), "Convergence Properties of a General Algorithm for Calculating Variational Bayesian Estimates for a Normal Mixture Model," *Bayesian Analysis*, 1, 625–650. [1149]

Wang, Y. J., and Wong, G. Y. (1987), "Stochastic Blockmodels for Directed Graphs," *Journal of the American Statistical Association*, 82, 8–19. [1155,1156]

Westling, T., and McCormick, T. H. (2015), "Beyond Prediction: A Framework for Inference with Variational Approximations in Mixture Models," arXiv:1510.08151. [1148,1149,1150,1155]

Yang, Y., Pati, D., and Bhattacharya, A. (2017), "α-Variational Inference With Statistical Guarantees," arXiv:1710.03266. [1149]

You, C., Ormerod, J. T., and Müller, S. (2014), "On Variational Bayes Estimation and Variational Information Criteria for Linear Regression Models," *Australian & New Zealand Journal of Statistics*, 56, 73–87. [1149]

Zhang, A. Y., and Zhou, H. H. (2017), "Theoretical and Computational Guarantees of Mean Field Variational Inference for Community Detection," arXiv:1710.11268. [1149]

Zhang, F., and Gao, C. (2017), "Convergence Rates of Variational Posterior Distributions," arXiv:1712.02519. [1149]