Journal of the Royal Statistical Society Series B: Statistical Methodology, 2024, **86**, 742–770
https://doi.org/10.1093/jrsssb/qkad164
Advance access publication 9 April 2024
Original Article

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series B
Statistical Methodology
B

Bayesian model selection via mean-field variational approximation

Yangfan Zhang and Yun Yang

Department of Statistics, University of Illinois Urbana-Champaign, Champaign, USA

Address for correspondence: Yun Yang, Department of Statistics, University of Illinois Urbana-Champaign, 605 E. Springfield Ave, Champaign, IL 61820, USA. Email: yy84@illinois.edu

Abstract

This article considers Bayesian model selection via mean-field (MF) variational approximation. Towards this goal, we study the non-asymptotic properties of MF inference that allows latent variables and model misspecification. Concretely, we show a Bernstein–von Mises (BvM) theorem for the variational distribution from MF under possible model misspecification, which implies the distributional convergence of MF variational approximation to a normal distribution centring at the maximal likelihood estimator. Motivated by the BvM theorem, we propose a model selection criterion using the evidence lower bound (ELBO), and demonstrate that the model selected by ELBO tends to asymptotically agree with the one selected by the commonly used Bayesian information criterion (BIC) as the sample size tends to infinity. Compared to BIC, ELBO tends to incur smaller approximation error to the log-marginal likelihood (a.k.a. model evidence) due to a better dimension dependence and full incorporation of the prior information. Moreover, we show the geometric convergence of the coordinate ascent variational inference algorithm, which provides a practical guidance on how many iterations one typically needs to run when approximating the ELBO. These findings demonstrate that variational inference is capable of providing a computationally efficient alternative to conventional approaches in tasks beyond obtaining point estimates.

Keywords: Bayesian inference, coordinate ascent, mean-field inference, oracle inequality

1 Introduction

Variational inference (VI, Bishop, 2006; Jordan et al., 1999) is an effective computational method for approximating complicated posterior distributions arising in Bayesian statistics. An alternative commonly adopted approach is the Markov Chain Monte Carlo (MCMC) method (Gelfand & Smith, 1990; Hammersley, 2013; Hastings, 1970) based on sampling, where a Markov chain is carefully constructed so that its limiting distribution matches the target distribution. Despite its popularity, MCMC is known to suffer from a number of drawbacks, including low-sampling efficiency due to sample correlation and a lack of computational scalability to massive datasets. In comparison, VI turns the integration or sampling problem into an optimization problem, and can be orders of magnitude faster than MCMC for achieving the same approximation accuracy. More precisely, the target distribution in VI is approximated by the closest member in a family of tractable distributions via minimizing the Kullback–Leibler (KL) divergence, which is also equivalent to maximizing a lower bound to the logarithm of the marginal density of data, called evidence lower bound (ELBO). We refer the interested readers to Blei et al. (2017) for a comprehensive review on the history of VI development.

Among various approximating schemes, the mean-field (MF) approximation, which uses the approximating family consisting of all fully factorized density functions over (blocks of) the target random variables, is the most widely used and representative instance of VI that is conceptually simple yet practically powerful. The computation of MF approximation can be realized using the coordinate ascent variational inference (CAVI) algorithm (Bishop, 2006; Blei et al., 2017),

Received: May 16, 2022. **Revised:** September 11, 2023. **Accepted:** December 8, 2023

© The Royal Statistical Society 2024. All rights reserved. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

which iteratively optimizes each (block of) components in the factorization while keeping others fixed at their present values (see Section 2.4 for more details). CAVI resembles the classical EM algorithm (Dempster et al., 1977) by viewing the target random variables as the unobserved latent variables. In this paper, we primarily focus on MF approximation, although our development can be analogously extended to other approximation schemes, such as Gaussian approximation and hybrid schemes that further restrict components in MF to be within certain exponential families.

Despite the wide applications of variational inference over the past decades, it is until recent years that some general theory, trying to explain from a frequentist perspective why variational inference works so well, has been developed. Some earlier threads of theoretical research are mostly conducted in a case-by-case manner, by either explicitly analysing the fixed point equation of the variational optimization problem, or directly analysing the iterative algorithm for solving the optimization problem. Examples of such include Bayesian linear models (Hall, Ormerod, et al., 2011; Hall, Pham, et al., 2011; Ormerod & Wand, 2012), Gaussian mixture models (GMMs) (Titterington & Wang, 2006; Westling & McCormick, 2015), and stochastic block models (SBMs) (Bickel et al., 2013; Zhang & Zhou, 2020). It is a well-known fact that many VI schemes, such as MF approximation, fail to capture the dependence structure and tend to underestimate the estimation uncertainty reflected in the Bayesian posteriors (Wang & Titterington, 2005). This observation is first theoretically illustrated by Wang and Blei (2019), where under a local asymptotic normality (LAN) assumption, a Bernstein–von Mises (BvM) type theorem (i.e. asymptotic normal approximation, Van der Vaart, 2000) is proven for the variational (approximated) posterior. However, the crucial LAN assumption used by Wang and Blei (2019) implicitly assumes the estimation consistency for the model parameter, which still requires a case-by-case verification. In addition, it is not clear that how the asymptotic mean and covariance matrix implied by their LAN assumption relate to characteristics of the statistical model in a general sense. Wang and Blei (2019) further generalized the results by Wang and Blei (2019) from well-specified models to misspecified models.

Since VI is generally incapable of accurately approximating the target posterior distributions, another line of research focuses on the estimation consistency of point estimators obtained from variational posteriors (e.g. using the expectation) under general settings. For example, a number of recent works (Alquier & Ridgway, 2020; Pati et al., 2018; Yang et al., 2020; Zhang & Gao, 2020) provide general conditions under which VI leads to consistent parameter estimation; moreover, they derive convergence rates of the point estimators that are often minimax-optimal (up to logarithmic terms) in both regular parametric and infinite-dimensional non-parametric models. In addition, some works (Alquier & Ridgway, 2020; Alquier et al., 2016; Chérief-Abdellatif & Alquier, 2018) derive risk bounds for VI under model misspecified settings; while Yang et al. (2020) also studied the risk bounds for VI with fractional posteriors. In a nutshell, this point estimation consistency (first-order information) can be attributed to the heavy penalty on the tails in the KL divergence that forces the variational distribution to concentrate around the true parameter at the optimal rate; however, the local shape of the obtained variational posteriors around the true parameter (second-order information) can be far away from that of the true posterior (see Figure 1 for an illustration).

Through a more refined analysis, Han and Yang (2019) showed that under some mildly stronger smoothness conditions without assuming consistency, the discrepancy between a point estimator from MF approximation and the maximum likelihood estimator (MLE) is of higher-order compared to the root-n estimation error of MLE for regular parametric models. As a consequence, there is essentially no loss of efficiency in using a VI point estimator for parameter estimation, in terms of asymptotically attaining the Cramér–Rao lower bound. In addition, they showed that MF variational posterior is close to a multivariate normal distribution centred at MLE with a diagonal precision matrix whose diagonal elements are equal to corresponding elements in the Fisher information at the true parameter. They further proposed a consistent variational weighted likelihood bootstrap method for uncertainty quantification for the MF approximation.

The overarching goal of the current paper is to carry out further methodological and theoretical investigations complimenting existing findings by looking into the model selection and algorithmic convergence aspects of MF variational approximation. While there are some existing theoretical results on model selection with variational inference, they either show only an oracle inequality for the selected model (Chérief-Abdellatif, 2019), implying that the estimation

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

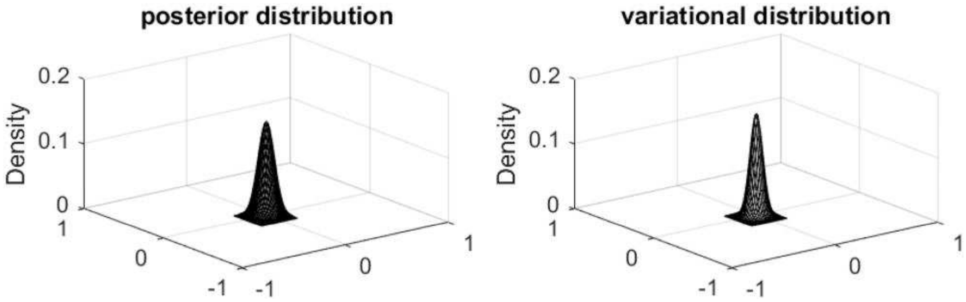

Figure 1. Plots of a two-dimensional posterior density and its mean-field approximation.

performance of the selected model is not much worse than that of the true model, or use an additional aggregation step to combine multiple variational models to achieve an optimal convergence rate (Ohn & Lin, 2021). It remains an open problem to study whether the model selected based on ELBO maximization is consistent; that is, whether the probability of choosing the right model tends to one as the sample size tends to infinity. In this paper, we propose ELBO as an alternative criterion for model selection, and show that it is asymptotically equivalent to the Bayesian information criterion (BIC, Schwarz, 1978). It is well-known (Yang, 2005) that BIC, derived as an asymptotic approximation to the log-marginal likelihood (a.k.a. model evidence), is consistent in selecting the true model; while another commonly used Akaike information criterion (AIC, Hirotugu, 1974), equivalent to Mallows's C_p statistic in regression analysis, is minimax-rate optimal for prediction but tends to overestimate the model size. In this work, we find that ELBO generally leads to a better approximation to the model evidence than BIC due to the full incorporation of prior information and a better dimension dependence in the approximation error. On the other hand, our numerical studies suggest that the proposed ELBO criterion tends to be less conservative than BIC in the presence of weak signals, and can achieve comparable predictive performance as AIC with a much more parsimonious selected model. Furthermore, we study the algorithmic convergence of the CAVI algorithm under regular parametric models by providing generic conditions on the initialization and step size under which CAVI exhibits geometric convergence towards the exact value up to the statistical accuracy of the problem. Specifically, we characterize the algorithmic convergence rate under two commonly adopted updating schemes, and our result provides theoretic guidance on how many iterations one typically needs to run when approximating the ELBO for model selection, so that the numerically selected model coincides with the theoretical one.

The rest of the paper is organized as follows. In Section 2, we review some background on the mean-field variational inference, present some new theoretical results on the misspecified Bernstein–von Mises theorem, and describe in detail the methodology of using MF approximation for model selection and the related computational aspects. Section 3 contains the main theoretical findings of the paper. Section 4 includes the numerical results. We conclude the paper with a discussion in Section 5 and leave a description of motivating examples, their theoretical consequences, high-dimensional extensions, and technical proofs to the online supplementary material, Appendices.

Notation. We use lower-case letters (e.g. π, p, q, …) to denote densities, and capital letters (e.g. Π, P, Q, …) for the associated probability measures. We use $D(P \parallel Q)$ to denote the KL divergence between two distributions P and Q, and $H(P, Q)$ for the Hellinger distance.

2 Model selection via mean-field approximation and its computation

In this section, we first set up the modelling framework and review MF variational approximation for Bayesian inference; and then present some new theoretical results, including concentration and distributional convergence under possible model misspecification. Motivated by the theory, we propose a new criterion based on the ELBO as an alternative to the widely used BIC for model selection. After that, we introduce several common variants of CAVI algorithms that implement MF for model selection. In online supplementary material, Appendix A, we provide some concrete motivating examples and derive the closed form of updating formulas for computing ELBO.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

2.1 Variational inference and mean-field approximation

Let $X^n = (X_1, \ldots, X_n)$ denote i.i.d. random observations from some unknown underlying data generating distribution $\mathbb{P}_0$ to be estimated. Consider a parametric family $\{\mathbb{P}_\theta : \theta \in \Theta\}$ that does not necessarily contain $\mathbb{P}_0$ (i.e. we allow model misspecification). In a general Bayesian framework, the goal is to approximate the posterior density π_n of parameter $\theta \in \Theta$ given data X^n, which is obtained by combining a prior density $\pi(\theta)$ and data likelihood $p(X^n \mid \theta)$,

$$\pi_n(\theta) := p(\theta \mid X^n) = \frac{p(X^n \mid \theta)\pi(\theta)}{p(X^n)}, \quad \text{with} \quad p(X^n) = \int_\Theta p(X^n \mid \theta)\pi(\theta)\,\mathrm{d}\theta$$

denoting the normalization constant. In many problems such as the Gaussian mixture modelling, model augmentation with latent variables can greatly simplify the likelihood function evaluation; thereby facilitating posterior calculation. For concreteness, we consider n local latent variables $S^n = (S_1, \ldots, S_n)$ where the ith latent variable $S_i \in \mathcal{S}$ is tied with X_i for $i = 1, 2, \ldots, n$. Under this setting, the marginal likelihood function of data X^n can be recovered by integrating the conditional density function $p(X^n \mid S^n, \theta)$ with respect to the latent variable distribution $p(S^n \mid \theta)$, i.e.

$$p(X^n \mid \theta) = \int_{\mathcal{S}^n} p(X^n \mid \theta, s^n) p(s^n \mid \theta)\,\mathrm{d}s^n.$$

In the case where latent variables are of discrete type, we replace the integration above with a summation over $\mathcal{S}^n$. We consider the common setting where the observation-latent variable pairs $\{(X_i, S_i)\}_{i=1}^n$ are conditionally i.i.d. given θ, i.e.

$$p(X^n \mid S^n, \theta) = \prod_{i=1}^n p(X_i \mid S_i, \theta) \quad \text{and} \quad p(S^n \mid \theta) = \prod_{i=1}^n p(S_i \mid \theta).$$

For such a latent variable model, one is typically interested in making inferences using the following joint posterior density over parameter θ and latent variables S^n,

$$p(\theta, S^n \mid X^n) = \frac{p(X^n \mid S^n, \theta) p(S^n \mid \theta) \pi(\theta)}{p(X^n)}.$$

Unfortunately, in many problems, the normalizing constant $p(X^n)$ involving a multivariate integration is analytically intractable and difficult to numerically approximate. VI instead searches for the closest distribution $\widehat{q}(\theta, S^n)$ over $\Theta \times \mathcal{S}^n$ from some computationally friendly variational family, denoted by Γ, to approximate the joint posterior distribution by solving the following optimization problem:

$$\widehat{q} = \operatorname*{argmin}_{q \in \Gamma} D\big(q(\cdot\,,\,\cdot) \,\|\, p(\cdot\,,\cdot \mid X^n)\big). \tag{1}$$

The following equivalent description of the objective functional above leads to an alternative way to interpret VI:

$$\begin{aligned} D\big(q(\cdot\,,\,\cdot) \,\|\, p(\cdot\,,\cdot \mid X^n)\big) &= \int_{\Theta \times \mathcal{S}^n} \log \frac{q(\theta, s^n)}{p(\theta, s^n \mid X^n)} q(\theta, s^n) \mathrm{d}\theta \mathrm{d}s^n \\ &= \int_{\Theta \times \mathcal{S}^n} \big[\log p(X^n) + \log q(\theta, s^n) - \log p(X^n, \theta, s^n)\big] q(\theta, s^n) \mathrm{d}\theta \mathrm{d}s^n := \log p(X^n) - L(q), \end{aligned} \tag{2}$$

where $L(q) = \int_{\Theta \times \mathcal{S}^n} [\log p(X^n, \theta, s^n) - \log q(\theta, s^n)] q(\theta, s^n) \mathrm{d}\theta \mathrm{d}s^n$ is called ELBO since it bounds the evidence $\log p(X^n)$ from below, due to the nonpositivity of KL divergence. The same identity also illustrates that VI circumvents the need to calculate the unknown normalizing constant $p(X^n)$ since it only contributes to the objective functional as a constant that does not change the optimum.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

According to identity (2), ELBO $L(q)$ equals to the evidence (i.e. $\log p(X^n)$) if and only if q and $p(\cdot, \cdot \mid X^n)$ coincide. As a direct consequence, the following two statistical tasks are equivalent: (1) approximating the joint posterior for conducting statistical inference on (θ, S^n) and (2) approximating the evidence $\log p(X^n)$ for evaluating model goodness of fit or performing model selection. Computationally, we can alternatively maximize the analytically tractable ELBO functional L to find the best approximation to the joint posterior within the variational family Γ. In this way, we have turned the integration problem into an optimization problem. Although VI is not guaranteed to generate exact samples as MCMC does, the computational efficiency can be considerably improved, as some smart choices of the variational family make the optimization problem numerically solvable and simple. Moreover, the computational speed can be further boosted by taking advantage of modern optimization techniques such as stochastic approximation and distributed computing.

In practice, a popular choice of Γ is the mean-field family that contains all fully factorized densities with the form $q(\theta, s^n) = \prod_{j=1}^{d} q_{\theta_j}(\theta_j) \cdot \prod_{i=1}^{n} q_{S_i}(s_i)$ for all $\theta \in \Theta$ and $s^n = (s_1, \ldots, s_n) \in \mathcal{S}^n$. As a representative illustrating example, the (closest) MF approximation to the multivariate normal distribution $N(\mu, \Theta^{-1})$ (with Θ being the precision matrix) is $N(\mu, [\mathrm{diag}(\Theta)]^{-1})$, where $\mathrm{diag}(A)$ denotes the diagonal matrix collecting all diagonal elements from a matrix A. In other words, the MF approximation keeps all the first-order information, while throwing away second-order interactions encoding the dependence structure, as we have already seen from Figure 1. A common variant relaxing the fully factorized MF approximation is the block MF approximation of the form $q(\theta, s^n) = q_\theta(\theta)q_{S^n}(s^n)$, where the dependence within components of θ or S^n is preserved while that between the two blocks θ and S^n is neglected. For either the full or the block MF approximation, the corresponding optimization problem can be efficiently solved by a generic coordinate ascent (Wright, 2015), whose specializations to MF will be introduced with more details in Section 2.4.

To interpret and understand different sources of errors due to MF approximation $q(\theta, s^n) = q_\theta(\theta)q_{S^n}(s^n)$, we may further decompose ELBO into the following three terms (Yang et al., 2020):

$$\begin{aligned} L(q) &= \int_{\Theta\times\mathcal{S}^n} \Big[\log\big\{p(s^n, X^n \mid \theta)\pi(\theta)\big\} - \log\big\{q_\theta(\theta)q_{S^n}(s^n)\big\}\Big] q_\theta(\theta)q_{S^n}(s^n)\mathrm{d}\theta \mathrm{d}s^n \\ &= \underbrace{\int_\Theta \log p(X^n \mid \theta)q_\theta(\theta)\mathrm{d}\theta - D(q_\theta \,\|\, \pi)}_{=:L_\theta(q_\theta)} - \Delta_J, \end{aligned} \tag{3}$$

where the first term is an integrated marginal log-likelihood function (w.r.t. the variational density q_θ), the second term is the negative KL divergence between q_θ and the prior, and the last term

$$\Delta_J = \int_\Theta \left[\log p(X^n \mid \theta) - \int_{\mathcal{S}^n} \log \frac{p(X^n, s^n \mid \theta)}{q_{S^n}(s^n)} q_{S^n}(s^n)\mathrm{d}s^n\right] q_\theta \mathrm{d}\theta \geq 0$$

can be interpreted as a non-negative Jensen gap introduced by approximating the marginal log-likelihood with a lower bound from Jensen's inequality. In particular, $L_\theta(q_\theta)$, which collects the first two terms in Eq. (3), constitutes the ELBO associated with the objective functional $D(q(\cdot) \,\|\, \pi_n(\cdot))$ in the variational inference of approximating the (marginal) posterior $p(\theta \mid X^n)$ with variational density $q_\theta(\theta)$; while the extra gap Δ_J is due to the presence of latent variables where $p(s^n \mid X^n, \theta)$ is approximated by a single distribution $q_{S^n}(s^n)$ for all $\theta \in \Theta$.

Decomposition (3) also reveals an interesting connection between the optimization of ELBO and the traditional regularized estimation. For example, when there is no latent variable, the Jensen gap term Δ_J vanishes, and maximizing the ELBO functional (3) over all density functions q_θ over Θ becomes finding a KL divergence-regularized estimator over Γ: the first term in Eq. (3) reflects model goodness of fit to the data and encourages q_θ to assign all its mass towards the maximizer of log-likelihood function, i.e. the maximum likelihood estimator; while the second regularization term $D(q_\theta \,\|\, \pi)$ avoids measure collapse as it diverges to infinity as q_θ becomes close to a point mass measure.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

2.2 Concentration and distributional convergence

This subsection presents results about large-sample properties of posterior distributions and their MF variational approximations in terms of concentration towards certain point in parameter space Θ and convergence towards some distribution over Θ, under the frequentist perspective assuming $\{X_i\}_{i=1}^n$ to be i.i.d. from the data generating distribution $\mathbb{P}_0$. These results are direct generalization of Han and Yang (2019) from well-specified parametric models to misspecified models. They also provide the cornerstone for showing the consistency of model selection based on ELBO in Section 3.1.

For a well-specified model where $\mathbb{P}_0 = \mathbb{P}_{\theta^*}$ for some true parameter θ^* in parameter space Θ, it is known that the posterior distribution tends to contract towards θ^* in view of the BvM theorem (Van der Vaart, 2000). For posterior concentration beyond parametric models, refer to Ghosal et al. (2000), Ghosal and Van Der Vaart (2007), and Shen and Wasserman (2001). However, in the context of model selection, we also need to consider candidate models that may not contain $\mathbb{P}_0$. Posterior contraction for misspecified models is formally studied by Kleijn and Van der Vaart (2012). More precisely, if we denote

$$\theta^*_{\mathcal{M}} = \operatorname*{argmin}_{\theta \in \Theta_{\mathcal{M}}} D(\mathbb{P}_0 \,\|\, \mathbb{P}_\theta) \tag{4}$$

as the parameter associated with the 'projection' (relative to the KL divergence) of $\mathbb{P}_0$ to a generic (possibly misspecified) model $\mathcal{M}$ with parameter space $\Theta_{\mathcal{M}}$. The BvM theorem under misspecification (Kleijn & Van der Vaart, 2012) states that the posterior distribution tends to be close to a normal distribution centring at MLE $\widehat{\theta}^{\text{mle}} = \arg\max_{\theta \in \theta_{\mathcal{M}}} p(X^n \mid \theta)$ of the assumed model $\mathcal{M}$, whose covariance matrix is related as usual to the Fisher information matrix; and the MLE $\widehat{\theta}^{\text{mle}}$ itself is asymptotically normal with asymptotic mean $\theta^*_{\mathcal{M}}$ and a sandwiched-form covariance matrix. Since we will fix the model $\mathcal{M}$ in the following analysis, we will omit the $\mathcal{M}$ in the subscripts in the rest of this subsection. For example, $\theta^*_{\mathcal{M}}$ will be simply denoted as θ^*, parameter space $\Theta_{\mathcal{M}}$ as Θ, and prior density $\pi_{\mathcal{M}}$ as π.

To begin with, we make some common regularity assumptions on the prior and log-likelihood function following Han and Yang (2019), with some appropriate generalization to the misspecified setting.

Assumption 1 (Thickness of prior). The prior satisfies $\pi(\theta^*) > 0$, and is continuously differentiable in a neighbourhood of θ^*, with growth controlled as

$$\left|\log \pi(\theta) - \log \pi(\theta^*)\right| \le C\left(1 + \|\theta - \theta^*\|^L\right),$$

for any θ and some positive constants (C, L).

Assumption 2 (Regularity of marginal likelihood). Let $\ell(\theta; x) = \log p(x \mid \theta)$ be the log-likelihood function at a single-data point x, then:

(a) $\ell(\theta; x)$ has continuous mixed derivatives up to order three with respect to components of θ, and the mixed derivatives have finite fourth moments in a neighbourhood of θ^* under $\mathbb{P}_0$. Moreover, there exists a measurable function $Z = Z(x)$ such that

$$\left|\frac{\partial^3 \ell(\theta; x)}{\partial\theta_i \partial\theta_j \partial\theta_k}\right| \le Z(x)\left(1 + \|\theta - \theta^*\|^L\right), \quad \text{for all } i, j, k \in [d],$$

and Z satisfies $\mathbb{E}_{\mathbb{P}_0}[e^{sZ(X)}] < \infty$ for some $s > 0$;

(b) The information matrix $V(\theta^*) := \mathbb{E}_{\mathbb{P}_0}[-\nabla^2 \ell(\theta^*, X)]$ is positive definite, and the covariance matrix $\mathbb{E}_{\mathbb{P}_0}[\nabla\ell(\theta^*; X)\nabla\ell(\theta^*; X)^T]$ of the score vector $\nabla\ell(\theta^*; X)$ is invertible;

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

(c) The Radon–Nikodym derivative $\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^*}}$ satisfies $\mathbb{E}_{\mathbb{P}_0}[\frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^*}}] < \infty$ in a neighbourhood of θ^*. Moreover, for any $\epsilon > 0$ there exists a sequence of test functions $\{\phi_n : n \geq 1\}$ such that

$$\mathbb{E}_{\mathbb{P}_0^n}[\phi_n(X^n)] \leq \exp(-nc) \quad \text{and}$$
$$\sup_{\|\theta-\theta^*\|\geq\epsilon} \mathbb{E}_{\mathbb{P}_0^n}\left[\frac{\mathbb{P}(X^n \mid \theta)}{\mathbb{P}(X^n \mid \theta^*)}\big(1 - \phi_n(X^n)\big)\right] \leq \exp(-nc_\epsilon),$$

for some constants c and c_ϵ.

Assumption 3 (Regularity of conditional likelihood of latent variable). The conditional log-likelihood function of the latent variable $\ell_S(\theta, X; s) = \log p(s \mid \theta, X)$ has continuous mixed derivatives with respect to components of θ up to order three, and the mixed derivatives have finite fourth moments in a neighbourhood of θ^* under $\mathbb{P}_0$. Moreover, there exists a measurable function $Z = Z(x, s)$ with bounded second moment under $\mathbb{P}_0$ such that

$$\left|\frac{\partial^3 \ell(\theta, x; s)}{\partial\theta_i \partial\theta_j \partial\theta_k}\right| \leq Z(x, s)\big(1 + \|\theta - \theta^*\|^L\big), \quad \text{for all } i, j, k \in [d].$$

Assumptions 1, 2(a), 2(b), and 3 are some standard regularity conditions made on the prior and log-likelihood functions for proving the BvM theorem (for well-specified models). Assumption 2(c) includes some additional regularity conditions and testability assumptions inherited from Kleijn and Van der Vaart (2012) to address misspecified models. We denote the latent variable Fisher information matrix (also called missing data Fisher information matrix in the missing data/EM algorithm literature, e.g. Dempster et al., 1977), $\mathbb{E}_{\mathbb{P}_0}[-\nabla^2 \ell_S(\theta^*, X)]$ of ℓ_S as $V_s(\theta^*)$ and denote $V_c(\theta^*) = V(\theta^*) + V_s(\theta^*)$ as the complete data information matrix at θ^*. In this work, we only consider nonsingular models by assuming the nonsingularity of the Fisher information (Assumption 2(b)). The problem of model selection involving possibly singular models (e.g. mixture models, Ho & Nguyen, 2019) is not covered by our theory and will be left to future research. As a result, in Section 4.2 for numerical study on Gaussian mixture model, we only consider a well-specified model. However, we note that our assumptions are satisfied for under-specified Gaussian mixture models, as well as for other commonly used models such as the linear and generalized linear models (GLM). In these cases, the Fisher information matrix remains nonsingular even when the model is misspecified, as discussed in Section 4.3.

Theorem 3.1 in Kleijn and Van der Vaart (2012) states that the posterior distribution of a misspecified model tends to be close to a normal distribution. In this paper, we find that the MF variational distribution also contracts to a normal distribution under model misspecification. Following some recently developed techniques, we first include the following two results on the consistency and concentration of the marginal posterior of θ and its MF approximation. The first result is a non-asymptotic version of Theorem 3.1 in Kleijn and Van der Vaart (2012) on the contraction of posterior under model misspecification, which is inherited from Theorem 5.1 in Ghosal et al. (2000).

Lemma 1 Under Assumptions 1 to 3, for any $K \geq 1$, it holds with probability at least $1 - CK^{-2}$ that the marginal posterior Π_n of θ satisfies

$$\Pi_n\big(\|\theta - \theta^*\| \geq C\varepsilon\big) \leq e^{-Cn\varepsilon^2}, \quad \text{for all } \varepsilon \geq K\varepsilon_n, \tag{5}$$

with $\epsilon_n = \frac{C\log n}{\sqrt{n}}$ for some constant C sufficiently large.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Using a similar variational argument as the proof of Lemma 2 in Han and Yang (2019), the sub-Gaussian type concentration property of the posterior $p(\theta \mid X^n)$ is inherited by its MF approximation $\widehat{Q}_\theta$, as described by the following lemma.

Lemma 2 Suppose the posterior satisfies the sub-Gaussian tail property as displayed in Lemma 1. Then under the same notation as Lemma 1, there exist constants (C_1, C_2, C_3) such that for any $K \geq 1$, it holds with probability at least $1 - C_1 K^{-2}$ that the MF approximation $\widehat{Q}_\theta$ satisfies

$$\widehat{Q}_\theta\big(\|\theta - \theta^*\| \geq C_2 \varepsilon\big) \leq e^{-C_3 n \varepsilon^2}, \quad \text{for all } \varepsilon \geq K \varepsilon_n.$$

Intuitively, the concentration of variational distribution $\widehat{Q}_\theta$ is a result of the strong penalty on the tail difference incurred by the log density ratio $\log \frac{q}{p}$ in the KL divergence objective (1), which forces $\widehat{Q}_\theta$ to concentrate around the same region where the posterior $p(\theta \mid X^n)$ assigns most of its mass. Lemma 2 further leads to the following theorem on the distributional convergence of $\widehat{Q}_\theta$, which generalizes Theorem 1 in Han and Yang (2019) from well-specified models to misspecified models. Note that the counterpart of Theorem 1 for well-specified models in Han and Yang (2019) is primarily used for studying the large-sample properties of the variational mean estimator $\widehat{\theta}_{VB} = \mathbb{E}_{\widehat{Q}_\theta}[\theta]$. In contrast, we extend and apply Theorem 1 to misspecified models to investigate the large-sample properties of the ELBO for model selection.

Theorem 1 Under Assumptions 1 to 3, there exist constants (C_4, C_5) and C^* that depend on the model and θ^* such that for any $1 \leq K = O(\sqrt{n})$ it holds with at probability at least $1 - C_4 K^{-2}$ that

$$D\big(\widehat{Q}_\theta \| Q^*_{VB}\big) \leq \frac{C_5 K^3 (\log n)^3}{\sqrt{n}},$$

where Q^*_{VB} denotes the normal distribution $N(\widehat{\theta}^{\text{mle}}, [n\, \text{diag}(V_c(\theta^*))]^{-1})$.

As a direct consequence of this theorem, any reasonable point estimator (e.g. by taking expectation) obtained from $\widehat{Q}_\theta$ is asymptotically the same as MLE $\widehat{\theta}^{\text{mle}}$ under the mis-specified model. The KL divergence bound from the theorem implies an $O(n^{-3/4})$ bound between the variational mean estimator $\widehat{\theta}_{VB} = \mathbb{E}_{\widehat{Q}_\theta}[\theta]$ and the MLE $\widehat{\theta}^{\text{mle}}$ by applying a transportation cost inequality. However, it is noteworthy that our analysis characterizing the updating dynamics of the CAVI iterative algorithm results in an improved error rate of $\|\widehat{\theta}_{VB} - \widehat{\theta}^{\text{mle}}\| = O(n^{-1}(\log n)^{9/2})$; see online supplementary material, Equation (39) in the proof of Theorem 4. This suggests that there is essentially no loss of statistical efficiency by using the MF approximation to obtain a point estimator to the model parameter as compared to the frequentist likelihood-based approaches in the misspecified setting. However, as is common in the MF variational inference, uncertainty quantification from $\widehat{Q}_\theta$ can be misleading as the asymptotic covariance matrix $[\text{diag}(V_c(\theta^*)]^{-1}$ may drastically underestimate the variability in the marginal posterior of θ, whose asymptotic covariance matrix is $[V(\theta^*)]^{-1}$ (Theorem 3.1, Kleijn & Van der Vaart, 2012). The uncertainty underestimation mainly comes from two sources: (1) ignoring the dependence among components of θ results in the diagonalization of the covariance matrix and (2) ignoring the dependence between parameter θ and latent variables S^n results in the inclusion of the extra $V_s(\theta^*)$ term in the precision matrix, i.e. the inverse of the covariance matrix. In particular, if there is no latent variable, then matrix V_c appearing in the Q^*_{VB} reduces to V.

Remark 1 (Extensions to Block MF). If we use the block mean-field approximation $q(\theta, s^n) = q_\theta(\theta) q_{S^n}(s^n)$, which preserves the within-block dependence as described in the previous subsection, instead of the fully factorized one as $q(\theta, s^n) = \prod_{j=1}^d q_{\theta_j}(\theta_j) \cdot \prod_{i=1}^n q_{S_i}(s_i)$, then Theorem 1 can be correspondingly

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

extended where the variational posterior Q^*_{VB} tends to be close to a normal with the same mean, but a different non-diagonal covariance matrix. Precisely, we consider the more general case of MF approximation using factorized densities with K blocks whose respective sizes are $d_1, \ldots, d_K$,

$$q_\theta(\theta) = q_1(\theta_1, \ldots, \theta_{d_1}) q_2(\theta_{d_1+1}, \ldots, \theta_{d_1+d_2}) \ldots q_K(\theta_{d-d_K+1}, \ldots, \theta_d).$$

We similarly write the complete data information matrix $V_c(\theta^*)$ in the corresponding block form as

$$V_c(\theta^*) = \begin{pmatrix} V_{11} & \cdots & V_{1K} \\ \vdots & \ddots & \vdots \\ V_{K1} & \cdots & V_{KK} \end{pmatrix},$$

where $V_{k\ell} \in \mathbb{R}^{d_k \times d_\ell}$ denotes the (k, ℓ)th block. Then $\widehat{Q}_\theta$ will be well-approximated by the normal distribution $N(\widehat{\theta}^{\text{mle}}, [nS_c(\theta^*)]^{-1})$ with

$$S_c(\theta^*) = \begin{pmatrix} V_{11} & 0 & \cdots & 0 \\ 0 & V_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & V_{KK} \end{pmatrix}.$$

In the special case of block MF over parameter θ and latent variables S^n, or $q(\theta, s^n) = q_\theta(\theta) q_{S^n}(s^n)$, the approximating normal distribution becomes $N(\widehat{\theta}^{\text{mle}}, (nV_c)^{-1})$, where the precision overestimation only comes from the extra latent variable information V_S in decomposition $V_c = V + V_S$. The profit regression example described in online supplementary material, Appendix A uses such a block approximation.

2.3 Bayesian model selection via mean-field approximation

In the previous subsection, we discussed the large-sample behaviour of MF variational inference in approximating the posterior. The normal approximation Q^*_{VB} to $\widehat{Q}_\theta$ suggests that we may use $\widehat{Q}_\theta$, or more precisely, the accompanied ELBO $L(\widehat{q}_\theta \otimes \widehat{q}_{S^n})$, as a computationally feasible surrogate to the evidence $\log p(X^n)$ for performing model selection. Concretely, recall that the true data generating distribution is denoted by $\mathbb{P}_0$. We consider a list of candidate models $\{\mathcal{M}_\lambda\}_{\lambda \in \Lambda}$ that may or may not contain $\mathbb{P}_0$. Our target is to select a most parsimonious model from $\{\mathcal{M}_\lambda\}_{\lambda \in \Lambda}$ that is closer to $\mathbb{P}_0$. In our setting of parameter models, we use $\theta_{\mathcal{M}_\lambda}$ and $\Theta_{\mathcal{M}_\lambda}$ to denote the respective parameter and parameter space associated with model $\mathcal{M}_\lambda$. The size (or complexity) of a model $\mathcal{M}_\lambda$ is then the dimension $d_{\mathcal{M}_\lambda}$ of $\Theta_{\mathcal{M}_\lambda}$ (i.e. number of parameters).

In the model selection literature, BIC (Schwarz, 1978) is a commonly used criterion function for selecting the best model that balances between goodness-of-fit to data and model complexity, defined as

$$\text{BIC}(\mathcal{M}) = -2\widehat{\ell}_n(\mathcal{M}) + d_{\mathcal{M}} \log n, \tag{6}$$

for a generic model $\mathcal{M}$ with $d_{\mathcal{M}}$ number of parameters, where $\widehat{\ell}_n(\mathcal{M}) = \max_{\theta_{\mathcal{M}} \in \Theta_{\mathcal{M}}} \log p(X^n \mid \theta_{\mathcal{M}}, \mathcal{M})$ denotes the maximal log-likelihood value under model $\mathcal{M}$. To employ BIC for model selection, one selects the model with the lowest BIC. More specifically, assume a prior distribution $\pi_{\mathcal{M}_\lambda}$ is imposed to $\theta_{\mathcal{M}_\lambda}$ under model $\mathcal{M}_\lambda$, and $p(\mathcal{M}_\lambda)$ denotes the prior probability assigned to model $\mathcal{M}_\lambda$. Using the Laplace approximation, it can be shown (Stoica & Selen, 2004, also see Theorem 2) that $-\text{BIC}(\mathcal{M}_\lambda)/2$ provides a large-sample approximation to the logarithm of the posterior probability $p(X^n \mid \mathcal{M}_\lambda)$ of data X^n given model $\mathcal{M}_\lambda$, i.e. the evidence of model $\mathcal{M}_\lambda$. Therefore, minimizing BIC over all candidate models is asymptotically equivalent to maximizing the posterior probability $p(\mathcal{M}_\lambda \mid X^n)$ over all $\lambda \in \Lambda$, as the impact from model prior $p(\mathcal{M}_\lambda)$ is diminishing as sample size n grows.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Since one perspective of variational inference as described in Section 2.1 is finding a best lower bound $\text{ELBO}(\mathcal{M}) = L_{\mathcal{M}}(\widehat{q}_{\mathcal{M}})$ to the evidence $\log p(X^n \mid \mathcal{M})$, where $L_{\mathcal{M}}$ denotes the ELBO under a generic model $\mathcal{M}$ and $\widehat{q}_{\mathcal{M}}$ denotes the MF approximation (i.e. solution of problem (1) that maximizes $L_{\mathcal{M}}$ within Γ), it is natural to use $\text{ELBO}(\mathcal{M})$ to approximate model evidence $\log p(X^n \mid \mathcal{M})$ to conduct model selection. Interestingly, we find (c.f. Section 3.1) that as the sample size n tends to infinity, model selection via maximizing the ELBO leads to the same model chosen via minimizing the BIC, which is also the highest posterior probability model. Since BIC is capable of consistently selecting the smallest model containing $\mathbb{P}_0$ as $n \to \infty$ (Schwarz, 1978), a property known as model selection consistency, we can conclude that ELBO inherits the same model selection consistency property.

Specifically, we find a precise characterization of the gap between the exact evidence $\log p(X^n \mid \mathcal{M})$ and the ELBO associated with MF approximation (Theorem 2) under a large-sample size n, which converges to a model-dependent constant $C^*(\mathcal{M}) = \frac{1}{2}\log\frac{\det(\text{diag}(V_c(\theta^*_{\mathcal{M}})))}{\det(V(\theta^*_{\mathcal{M}}))}$ as $n \to \infty$. In comparison, $-\text{BIC}(\mathcal{M})/2$ also provides an asymptotically constant approximation to the evidence, where the limiting constant is $C^*_{\text{BIC}}(\mathcal{M}) = -\frac{1}{2}\log\det(V(\theta^*_{\mathcal{M}})) + \frac{d_{\mathcal{M}}}{2}\log(2\pi) + \log\pi_{\mathcal{M}}(\theta^*_{\mathcal{M}})$ (Theorem 2). The empirical results in Section 4 also align well with these theoretical findings. Consequently, approximating the evidence by ELBO does not incur significantly larger error than that by BIC. Moreover, since the discrepancy between the negative half of BIC and ELBO is of constant order while the optimality gap, i.e. the smallest difference in the BIC values between the optimal model and any suboptimal model will be at least of order $\log n$ in order for the true model to be statistically identifiable, they tend to lead to the same selected model as $n \to \infty$.

More interestingly, a closer inspection of the two limiting constants $C^*(\mathcal{M})$ and $C^*_{\text{BIC}}(\mathcal{M})$ reveals that ELBO generally leads to a better approximation of the model evidence than BIC due to the full incorporation of prior information and a better dimension dependence in the approximation error. For example, the definition (6) of BIC completely ignores the prior contribution while ELBO (defined after Eq. (2)) involves an integration of log-prior, explaining the extra $\log\pi_{\mathcal{M}}(\theta^*_{\mathcal{M}})$ term in the BIC approximation error. In addition, due to the simple Laplace approximation, $C^*_{\text{BIC}}(\mathcal{M})$ has an extra term explicitly dependent of model size $d_{\mathcal{M}}$, making BIC less accurate in approximating the evidence in large or growing dimension problems; also see our numerical study in Section 4.3 on variable selection in GLM. Last but not least, our numerical studies in Section 4 suggest that ELBO tends to be less conservative than BIC in the presence of weak signals, and can achieve comparable predictive performance as AIC (Hirotugu, 1974) with a much more parsimonious selected model. Note that AIC is known to be minimax-rate optimal for prediction in linear regression but tends to overestimate the model size (Yang, 2005).

We conclude this section with a brief discussion on computation. Variational inference is commonly used when the posterior does not have an explicit form (e.g. in mixture models). In such settings, the MLE required for BIC calculation also does not admit a closed-form solution. In these cases, the MLE can be solved numerically using the EM algorithm or approximated by the expectation of the variational distribution, where the latter incurs an error of order at most $O(n^{-1})$ (see the remark after Theorem 1). In either scenario, the computational cost for calculating the MLE (or BIC) is comparable to that of variational inference or CAVI, since the EM algorithm can be viewed as a degenerate CAVI in which the parameter component in the mean-field approximation is further restricted to be a point mass measure. Therefore, previous discussions (or Theorem 2) suggest that ELBO tends to achieve better approximation accuracy than BIC with similar computational costs when estimating model evidence. Additionally, a standard implementation of CAVI for computing the variational distribution requires $O(n)$ computational cost per iteration due to the necessity of accessing the full data set. In settings where n is large, we may consider approximation methods such as stochastic gradient descent to reduce the per iteration computational cost (Alquier & Ridgway, 2020; Titsias & Lázaro-Gredilla, 2014). We would also like to note that calculating the BIC often requires accurately determining the effective dimension of the model. For some complex models with non-trivial constraints, such as Bayesian factor models (with constraints on the factor loading matrix to enforce identifiability) or latent variable models like Hidden Markov models, counting this effective dimension might not be straightforward. In contrast, calculating the ELBO is a straightforward by-product of implementing the MF variational inference, which automatically incorporates the effective dimension and eliminates the need for case-by-case analysis.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

2.4 Computation via coordinate ascent

Coordinate ascent is a natural and efficient algorithm for optimizing over densities taking a product form as in MF approximation (1). For illustration, we focus on models without latent variables, where the mean-field family contains all factorized densities of the form $q(\theta) = q_1(\theta_1)q_2(\theta_2)\cdots q_d(\theta_d)$ for $\theta \in \mathbb{R}^d$. Otherwise, we may view the latent variables as one (block) coordinate of the parameter and derive the corresponding coordinate ascent algorithms. In general, we may also block approximation where each component is a multi-dimensional density. The key idea of coordinate ascent is to optimize over one component of $q(\theta)$ at a time while fixing the others. Let $q_{-j}(\theta_{-j}) = \prod_{\ell \neq j} q_\ell(\theta_\ell)$ denote the joint density function of θ_{-j}, all components in θ except for θ_j. When optimizing over the jth component q_j, it would be helpful to express ELBO as a functional of q_j,

$$\begin{aligned} L(q_j; q_{-j}) &= \int_{\mathbb{R}^d} q(\theta)\big\{\log p(X^n, \theta) - \log q(\theta)\big\}\,\mathrm{d}\theta \\ &= \int_{\mathbb{R}} q_j(\theta_j)\left[\int_{\mathbb{R}^{d-1}} q_{-j}(\theta_{-j}) \log p\big(\theta_j \mid \theta_{-j}, X^n\big)\,\mathrm{d}\theta_{-j}\right]\mathrm{d}\theta_j - \int_{\mathbb{R}} q_j(\theta_j) \log q_j(\theta_j)\,\mathrm{d}\theta_j + C(\theta_{-j}), \end{aligned}$$

where constant $C(\theta_{-j})$ is independent of θ_j. From this identify, we may explicitly solve the optimizer $q_j^*(\theta_j) := arg\,max_{q_j} L(q_j; q_{-j})$ as

$$q_j^*(\theta_j) \propto \exp\left\{\int_{\mathbb{R}^{d-1}} q_{-j}(\theta_{-j}) \log p\big(\theta_j \mid \theta_{-j}, X^n\big)\,\mathrm{d}\theta_{-j}\right\}. \tag{7}$$

As is often in practice, one can recognize q_j^* above as coming from some parametric family, for example, certain exponential family, and determine the normalizing constant of q_j^*; otherwise, either particle methods (Saeedi et al., 2017; Wang et al., 2021) can be employed to approximate q_j^* over $\mathbb{R}$, or q_j can be further restricted to some parametric family, resulting in a hybrid variational approximation. To summarize, we can apply coordinate ascent to numerically solve the optimization problem in MF based on Eq. (7) in an iterative manner; and the resulting method is known as the CAVI (Bishop, 2006) in the literature. To avoid overly aggressive moves that may lead to periodic oscillation or even divergence, it is customary to introduce a step size parameter $\gamma \in (0, 1]$ and define the next iterate as proportional to the weighted geometric average $(q_j^*)^\gamma q_j^{1-\gamma}$ between q_j^* and current iterate q_j. In particular, with full step size $\gamma = 1$, the update is most greedy and the coordinate ascent becomes alternating maximization. As we will show in our theoretical analysis in Section 3.2 and some numerical studies in Section 4, the introduction of a partial step size $\gamma \in (0, 1)$ is necessary in some problems to avoid algorithmic nonconvergence.

In practice, two common updating schemes are utilized in CAVI, depending on whether the components are sequentially or simultaneously updated. In order to draw a connection with two well-known iterative algorithms for solving linear systems (e.g. Trefethen & Bau III, 1997), we will exchangeably call the followings as the Jacobi scheme and Gauss–Seidel scheme respectively, at iteration $t = 0, 1, \ldots$:

Parallel (Jacobi) update. We update all d components simultaneously based on the last iteration. To be more precise, we run following d updates at each iteration:

$$q_j^{(t+1)} \propto \exp\left\{\int_{\mathbb{R}^{d-1}} q_{-j}^{(t)}(\theta_{-j}) \log p\big(\theta_j \mid \theta_{-j}, X^n\big)\,\mathrm{d}\theta_{-j}\right\}, \quad j = 1, \ldots, d. \tag{8}$$

This algorithm can be run in parallel. However, the ELBO is not guaranteed to be monotonically decreasing in t if the full step size is used. To guarantee the convergence, we may need to take a step

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

size $\gamma \in (0, 1]$, i.e.

$$q_j^{(t+1)} \propto \exp\left\{\gamma \int_{\mathbb{R}^{d-1}} q_{-j}^{(t)}(\theta_{-j})\big[\log p(\cdot \mid \theta_{-j}, X^n)\big] d\theta_{-j}\right\}[q_j^{(t)}]^{1-\gamma}, \quad j = 1, \ldots, d. \tag{9}$$

This however may lead to a slower convergence.

Randomized sequential (Gauss–Seidel) update. We update one component at each time using the most recent updates of the other components. For simplicity, we focus on the randomized sequential update, where a component is randomly picked to be updated. Concretely, for a random index $j(t) \sim \text{Unif}(1, \ldots, d)$, we compute

$$q_{j(t)}^{(t+1)} \propto \exp\left\{\int_{\mathbb{R}^{d-1}} q_{-j(t)}^{(t)}\big(\theta_{-j(t)}\big) \log p\big(\theta_{j(t)} \mid \theta_{-j(t)}, X^n\big)\, \mathrm{d}\theta_{-j(t)}\right\}. \tag{10}$$

In this way, ELBO value is guaranteed to be monotonically nondecreasing in t. However, due to the sequential nature, we cannot utilize parallel computation techniques to reduce the run time. For the sequential update, we may also consider the systematic variant where the components are updated in a prespecified deterministic order.

Due to the numerical error, the computed ELBO value may differ from the theoretical ELBO value. It is therefore necessary to analyse the convergence of CAVI to guide its implementation in practice, particularly when applied to perform model selection. In Section 3.2, we determine the algorithmic convergence rate of CAVI, and study the impact of various problem characteristics on the rate. Most literature in CAVI convergence studies some special cases such as Gaussian mixture models (Titterington & Wang, 2006; Wang & Titterington, 2005), stochastic block models (Mukherjee et al., 2018; Sarkar et al., 2021; Yin et al., 2020; Zhang & Zhou, 2020), and Ising models (Jain et al., 2018; Koehler, 2019; Plummer et al., 2020). Accessing the convergence and determining the accompanied rate under general settings is still an open problem.

In Section 3.2, we identify a set of suitable conditions under which the tth iteration $q^{(t)}$ from CAVI satisfies

$$\mathbb{E}\big[L(\widehat{q}) - L(q^{(t)})\big] \le \alpha^t \mathbb{E}\big[L(\widehat{q}) - L(q^{(0)})\big] + \delta_n$$

as long as initialization $q^{(0)}$ is in a constant neighbourhood around the true parameter θ^*, where α is the algorithmic convergence rate depending on the model and the CAVI setup, and $\delta_n = \mathcal{O}((\log n)^3/\sqrt{n})$ is the usual root-n statistical error, where the poly-log n factor in δ_n is due to the high-probability argument in our proof. This result can be interpreted as that under a warm initialization, the ELBO regret (the difference between the ELBO value $L(\widehat{q})$ at the optimum $\widehat{q}$ and $L(q^{(t)})$) relative to the theoretical value has a geometric convergence towards zero up to a statistical error of the problem. Since $\mathbb{E}[L(\widehat{q}) - L(q^{(0)})]$ has a trivial bound as $\mathcal{O}(nd)$, to guarantee the output ELBO value to be within n^{-c} distance away from its theoretical value, our theory suggests that $O(d \log(nd))$ iterations suffice (so that each component is updated $O(\log(nd))$ times on average). Moreover, under a suitable metric, which will be the Hellinger distance $H(\cdot, \cdot)$, we show that the iterate $q^{(t)}$ at time t from CAVI converges towards $\widehat{q}$ at a geometric speed up to a statistical error term δ'_n as

$$\mathbb{E}[H(q^{(t)}, \widehat{q})] \le C_0 \alpha^{t/2} + \delta'_n,$$

where the convergence rate is expected to be $\sqrt{\alpha}$ since the KL divergence is locally quadratic when expanded in $H(q^{(t)}, \widehat{q})$. This result can be used to determine the number of iterations for obtaining a root-n consistent point estimator based on the output of q_θ from CAVI.

Another interesting finding implied by our theory is that the randomized sequential update always converges exponentially fast with full step size given a warm initialization. In contrast, in some examples with moderate dependence, the parallel update will converge only when a partial step size (i.e. $\gamma \in (0, 1)$) is used. However, for those step sizes under which both schemes converge, they tend to exhibit a similar convergence behaviour.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

3 Theoretical results

In this section, we provide the main theoretical results of this paper. In Section 3.1, we derive a non-asymptotic expansion of ELBO under MF approximation, and compare it with that of the BIC. We also build an oracle inequality, showing that the model selected by ELBO has the prediction performance comparable to the best model, even in the case all candidate models are misspecified and do not contain $\mathbb{P}_0$. In Section 3.2, we analyse the algorithmic convergence of the CAVI under the two aforementioned schemes, and discuss their consequences.

3.1 Model selection consistency based on ELBO

In this subsection, we show the consistency of model selection based on (penalized) ELBO. Our first result provides non-asymptotic expansions of ELBO and BIC for approximating the model evidence $\log p(X^n \mid \mathcal{M})$. Recall that $\theta^*_{\mathcal{M}}$ defined in Eq. (4) denotes the parameter in model $\mathcal{M}$ such that $\mathbb{P}_{\theta^*_{\mathcal{M}}}$ best approximates the true data generating distribution $\mathbb{P}_0$.

Theorem 2 Suppose the same assumptions of Theorem 1 to hold for a generic model $\mathcal{M}$. There exists constants C_6, C_7 such that for any $1 \le K = O(\sqrt{n})$, it holds with probability at least $1 - C_6 K^{-2}$ that

$$\left|\mathrm{ELBO}(\mathcal{M}) - \log p(X^n \mid \mathcal{M}) + C^*(\mathcal{M})\right| \le \frac{C_7 K^3 (\log n)^3}{\sqrt{n}}, \tag{11}$$

$$\left|-\frac{1}{2}\mathrm{BIC}(\mathcal{M}) - \log p(X^n \mid \mathcal{M}) + C^*_{\mathrm{BIC}}(\mathcal{M})\right| \le \frac{C_7 K^3 (\log n)^3}{\sqrt{n}}, \tag{12}$$

where $C^*(\mathcal{M}) = \frac{1}{2}\log\frac{\det(\mathrm{diag}(V_c(\theta^*_{\mathcal{M}})))}{\det(V(\theta^*_{\mathcal{M}}))}$, and

$$C^*_{\mathrm{BIC}}(\mathcal{M}) = -\frac{1}{2}\log\det(V(\theta^*_{\mathcal{M}})) + \frac{d_{\mathcal{M}}}{2}\log(2\pi) + \log \pi_{\mathcal{M}}(\theta^*_{\mathcal{M}}).$$

Note that inequalities (11) and (12) together imply the difference between $-\mathrm{BIC}/2$ and ELBO to be asymptotically constant as $n \to \infty$, which we denote as

$$\widetilde{C}^*(\mathcal{M}) \triangleq C^*(\mathcal{M}) - C^*_{\mathrm{BIC}}(\mathcal{M}) = \frac{1}{2}\log\det(\mathrm{diag}(V_c(\theta^*_{\mathcal{M}}))) - \frac{d_{\mathcal{M}}}{2}\log(2\pi) - \log\pi_{\mathcal{M}}(\theta^*_{\mathcal{M}}). \tag{13}$$

As a direct consequence of the theorem, the model selected by maximizing ELBO will lead to the same one that minimizes BIC, as long as the minimal BIC value is at least of order $\log n$ away from the second smallest value. For example, let $\mathcal{M}^*$ denote the smallest model containing the data generating distribution $\mathbb{P}_0$. Under the model identifiability condition that any underfitted model $\mathcal{M}$ not containing $\mathbb{P}_0$ has a strictly positive 'KL gap' $\inf_{\theta\in\Theta_{\mathcal{M}}} D(\mathbb{P}_0 \,\|\, \mathbb{P}_\theta) \ge \varepsilon > 0$ for some $\varepsilon > 0$, it is not difficult to show (e.g. see online supplementary material, Appendix D.5) that the difference $\mathrm{BIC}(\mathcal{M}) - \mathrm{BIC}(\mathcal{M}^*)$ between BIC values of $\mathcal{M}$ and $\mathcal{M}^*$ is at least $n\varepsilon - d_{\mathcal{M}^*}\log n \gg \log n$ with high probability. On the other hand, for any overfitted model $\mathcal{M}$ containing $\mathbb{P}_0$ and having a larger number of parameters $d_{\mathcal{M}} \ge d_{\mathcal{M}^*} + 1$, we have that $|\widehat{\ell}_n(\mathcal{M}) - \widehat{\ell}_n(\mathcal{M}^*)| = o(\log n)$ holds with high probability (see online supplementary material, Appendix D.5); this further implies $\mathrm{BIC}(\mathcal{M}) - \mathrm{BIC}(\mathcal{M}^*) \ge (d_{\mathcal{M}} - d_{\mathcal{M}^*})\log n - o(\log n) \gtrsim \log n$. By combining two cases, we can conclude that maximizing ELBO will consistently select the best model $\mathcal{M}^*$ as by minimizing the BIC since $|-\mathrm{BIC}(\mathcal{M})/2 - \mathrm{ELBO}(\mathcal{M})| \le C$ for some constant $C > 0$ when n is sufficiently large.

Note that the same discussion in Remark 1 also applies to Theorem 2 for the block mean-field approximation, where the constant then becomes $C^*_{\mathrm{b}}(\mathcal{M}) = \frac{1}{2}\log\frac{\det(S_c(\theta^*_{\mathcal{M}}))}{\det(V(\theta^*_{\mathcal{M}}))}$, which is never larger than the $C^*(\mathcal{M})$ in Theorem 2. This observation suggests that incorporating additional structure in the VI is always beneficial for improving the approximation accuracy given the computation is still tractable. In the other situation with no latent variables, the constant becomes

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

$C_{\mathrm{n}}^*(\mathcal{M}) = \frac{1}{2}\log\frac{\det(\mathrm{diag}(V(\theta^*_{\mathcal{M}})))}{\det(V(\theta^*_{\mathcal{M}}))}$, which is precisely the KL divergence between the two normal distributions $N(\widehat{\theta}^{\mathrm{mle}}, [nV(\theta^*_{\mathcal{M}})]^{-1})$ and $N(\widehat{\theta}^{\mathrm{mle}}, [n\,\mathrm{diag}(V(\theta^*_{\mathcal{M}}))]^{-1})$, where the former approximates the posterior distribution under model misspecification (Kleijn & Van der Vaart, 2012) and the latter approximates the MF solution $\widehat{Q}_\theta$ (Theorem 1 without latent variables), respectively. This limiting gap $C_{\mathrm{n}}^*(\mathcal{M})$ will vanish if components of $\widehat{\theta}^{\mathrm{mle}}$ are asymptotically independent (e.g. the location-scale normal model in Section 4.1).

Remark 2 (Discussion on model selection criteria). For a Bayesian model with latent variables, we have another ELBO value to use for model selection, which is the parameter part $L_{\theta_{\mathcal{M}}}(q_{\theta_{\mathcal{M}}})$ in decomposition (3), or

$$\begin{aligned} L_{\theta_{\mathcal{M}}}(q_{\theta_{\mathcal{M}}}) = L(q_{\mathcal{M}}) + \Delta_J(q_{\mathcal{M}}) &= \int_{\Theta_{\mathcal{M}}} \log\frac{p(X^n, \theta_{\mathcal{M}})}{q_{\theta_{\mathcal{M}}}(\theta_{\mathcal{M}})} q_{\theta_{\mathcal{M}}}(\theta_{\mathcal{M}})\,\mathrm{d}\theta_{\mathcal{M}} \\ &= \log p(X^n \mid \mathcal{M}) - D\big(q_{\theta_{\mathcal{M}}} \,\big|\big|\, \pi_n\big), \end{aligned}$$

where recall that π_n denotes the marginal posterior of θ and the extra Jensen gap $\Delta_J(q_{\mathcal{M}})$ is due to latent variables. In the proof of Theorem 2, it is shown that this Jensen gap asymptotically converges to a nonnegative constant $\frac{1}{2}\mathrm{tr}([\mathrm{diag}(V_c)]^{-1}V_s)$. Therefore, the parameter part $\mathrm{ELBO}_\theta(\mathcal{M}) := L_{\theta_{\mathcal{M}}}(\widehat{q}_{\theta_{\mathcal{M}}})$ leads to an improved approximation to the evidence (due to a smaller gap); and model selections based on ELBO_θ and ELBO are asymptotically equivalent, and equivalent to that based on BIC. However, unlike ELBO that can be efficiently computed via CAVI, ELBO_θ may not admit a closed form updating formula; and requires Monte Carlo methods to approximate.

Bayesian hypothesis testing can be viewed as a special instance of model selection with two candidate models, denoted as $\mathcal{M}_0$ and $\mathcal{M}_1$. Decisions in Bayesian hypothesis testing or model selection between two models are typically based on the so-called Bayes factor, which is defined as

$$B(\mathcal{M}_0, \mathcal{M}_1) = \frac{p(X^n \mid \mathcal{M}_0)}{p(X^n \mid \mathcal{M}_1)}.$$

Motivated by our general model selection procedure based on ELBO, we define the ELBO factor $\mathrm{ELBO}(\mathcal{M}_0) - \mathrm{ELBO}(\mathcal{M}_1)$ as a computationally-efficient surrogate to the log-Bayes factor, by noticing

$$\log B(\mathcal{M}_0, \mathcal{M}_1) = \log\frac{P(X^n \mid \mathcal{M}_0)}{P(X^n \mid \mathcal{M}_1)} = \mathrm{ELBO}(\mathcal{M}_0) - \mathrm{ELBO}(\mathcal{M}_1) + \mathcal{O}_p(1),$$

where the second equality is due to Theorem 2. We summarize the result in the following.

Corollary 1 Suppose the assumptions of Theorem 1 hold for the two model candidates $\mathcal{M}_0$ and $\mathcal{M}_1$. Then we may approximate the Bayes factor $-\frac{1}{2}B(\mathcal{M}_0, \mathcal{M}_1) := -\frac{1}{2}\frac{P(X^n|\mathcal{M}_0)}{P(X^n|\mathcal{M}_1)}$ based on the ELBO, as we have with probability at least $1 - C_6K^{-2}$ that

$$\begin{aligned} &\left| -\frac{\log B(\mathcal{M}_0, \mathcal{M}_1)}{2} - [\mathrm{ELBO}(\mathcal{M}_0) - \mathrm{ELBO}(\mathcal{M}_1)] - [\widetilde{C}^*(\mathcal{M}_0) - \widetilde{C}^*(\mathcal{M}_1)] \right| \\ &\leq \frac{C_7K^5(\log n)^3}{\sqrt{n}}, \end{aligned}$$

for some constants C_6, C_7, and constants $\widetilde{C}^*(M_i)$ ($i = 0, 1$) are given by Eq. (13).

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Previously, we have seen that model selection based on ELBO tends to agree with that based on BIC when at least one of the candidate models contains $\mathbb{P}_0$ and it is statistically distinguishable. However, it is often in practice that none of the models in the list contains $\mathbb{P}_0$. Another common situation is when some signals are weak, for example, the β_{min}-condition (c.f. Yang et al., 2016) does not hold in linear regression with moderate to large number of variables; then it is information-theoretically impossible to select all important variables. Moreover, using a smaller model may lead to better prediction performance than the true model by excluding those less significant variables; in fact, using the true model with many variables may lead to overfitting. For example, in the numerical study of variable selection in GLM in Section 4.3, we can see from the left two plots in Figure 8 that both BIC and ELBO tend to select a smaller model which turns out to have better prediction performance than the true model with five variables. To theoretically quantify the quality of the model selected via ELBO in the MF approximation in these situations, we prove an oracle inequality, which shows that the model selected by ELBO has prediction performance comparable to the best model in the list, even all candidate models may be misspecified and do not contain $\mathcal{P}_0$.

Theorem 3 Under the same assumptions as Theorem 2, there exists some constant $C_0 > 0$ such that for any $1 \le K = O(\sqrt{n})$, it holds with probability at least $1 - CK^{-2}$ that the model selected by ELBO $\widehat{\mathcal{M}} = arg\,max_{\mathcal{M}}\text{ELBO}(\mathcal{M})$ satisfies

$$\int_{\Theta_{\widehat{\mathcal{M}}}} D\big(\mathbb{P}_0||\mathbb{P}_{\theta_{\widehat{\mathcal{M}}}}\big)\widehat{q}_{\theta_{\widehat{\mathcal{M}}}}(\theta_{\widehat{\mathcal{M}}})\,\mathrm{d}\theta_{\widehat{\mathcal{M}}} \le \inf_{\mathcal{M}}\left\{\inf_{\theta_{\mathcal{M}}\in\Theta_{\mathcal{M}}} D\big(\mathbb{P}_0||\mathbb{P}_{\theta_{\mathcal{M}}}\big) + \frac{d_{\mathcal{M}}\log n}{2n}\right\} + \frac{C_0}{n} + \frac{CR^3K^3(\log n)^3}{\sqrt{n}},$$

where R is the total number of candidate models.

As mentioned in Section 1, Chérief-Abdellatif (2019) also proved an oracle inequality for the model selected by ELBO maximization in variational inference. However, their results apply only to models that do not contain latent variables, and their risk function is based on the α-Rényi divergence, which is generally weaker than the KL divergence used in our results. Additionally, their result is stated only for α-fractional posteriors under $\alpha < 1$, which requires fewer assumptions than our Assumptions 2 and 3. Specifically, they do not require a testing condition such as Assumption 2(c), which is one merit of considering Bayesian fractional posteriors (Bhattacharya et al., 2019).

3.2 Convergence of CAVI algorithms

In this subsection, we address the theoretical question of analysing the convergence of CAVI algorithms. The algorithmic convergence analysis provides a theoretical guidance on how many iterations are required to adequately approximate ELBO to achieve model selection consistency; according to Theorem 2 and the related discussions, a constant numerical error approximation to the ELBO would suffice for this purpose. Since in the convergence analysis the model $\mathcal{M}$ is fixed throughout, we will omit all $\mathcal{M}$ in the subscripts when no ambiguity may arise. We let $D^{(t)} = D(q^{(t)} \,\|\, \pi_n) - D(\widehat{q} \,\|\, \pi_n)$ be a discrepancy measure between the tth iterate $q^{(t)}$ from CAVI and the MF solution $\widehat{q}$. Note that we also have $D^{(t)} = L(\widehat{q}) - L(q^{(t)})$, which is the regret of $q^{(t)}$ relative to the theoretical ELBO value achieved by $\widehat{q}$.

Special case: Gaussian posterior without latent variables. We first consider the special case of a Gaussian posterior to help explain the intuition. Specifically, we assume the target posterior π_n to be the density of $N(\widehat{\theta}^{\text{mle}}_{\mathcal{M}}, [nV(\theta^*_{\mathcal{M}})]^{-1})$, which will be denoted by ϕ_n. Since our later analysis concerning a general posterior π_n will have a leading term related to its (asymptotic) normal approximation $N(\widehat{\theta}^{\text{mle}}_{\mathcal{M}}, [nV(\theta^*_{\mathcal{M}})]^{-1})$, we will preserve the notation $p^{(t)}(\theta_1, \ldots, \theta_d) = \prod_{j=1}^{d} p_j^{(t)}(\theta_j)$ for the tth iterate in the CAVI, either the parallel or the randomized sequential update depending on the context, of maximizing $D(p \,\|\, \phi_n)$ in the MF variational inference. To simplify the notation, we use the shorthand notation $\widehat{\theta} = \widehat{\theta}^{\text{mle}}_{\mathcal{M}}$, $V = V(\theta^*_{\mathcal{M}})$ and $S = \text{diag}(V)$. Under this notation, the MF approximation is $\phi^* := \arg\min_{p=\otimes_{j=1}^{d} p_j} D(p \,\|\, \phi_n)$, which is the density of $N(\widehat{\theta}, (nS)^{-1})$.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

To simplify the analysis for the Gaussian posterior, let us assume the initialization $p^{(0)}$ to be the density of $N(\theta^{(0)}, (nS)^{-1})$ for some initial mean vector $\theta^{(0)} \in \mathbb{R}^d$. This assumption is not overly restrictive. It is easy to verify that after each component in $p = \otimes_{j=1}^{d} p_j$ is updated at least once with full step size $\gamma = 1$, all future iterates of p will follow a normal distribution with covariance matrix $(nS)^{-1}$. Due to this reason, let us denote $p^{(t)}$ as the density of $N(\theta^{(t)}, (nS)^{-1})$. To describe the two updating schemes of the CAVI described in Section 2.4, let us derive the common step of updating one component p_ℓ from the following rule. Recall that the parallel scheme updates all components with indices $\ell = 1, 2, \ldots, d$ in one iteration; while the randomized sequential scheme randomly picks one index ℓ. The common updating formula for p_ℓ is

$$p_\ell^{(t+1)}(\theta_\ell) \propto \exp\Big(-\frac{\gamma n}{2}\mathbb{E}_{p_{-\ell}^{(t)}}\big[(\theta - \widehat{\theta})^T V(\theta - \widehat{\theta})\big]\Big) \cdot \big[p_\ell^{(t)}(\theta_\ell)\big]^{1-\gamma},$$

where $\gamma \in (0, 1]$ denotes the step size. Denote the bias at iteration t as $b^{(t)} = \theta^{(t)} - \widehat{\theta}$. Then it is straightforward to translate the preceding display into an updating rule for $b^{(t)}$ as

$$b_\ell^{(t+1)} = (1 - \gamma) b_\ell^{(t)} - \gamma \sum_{k \neq \ell} \frac{V_{\ell k}}{V_{\ell\ell}} b_k^{(t)}. \tag{14}$$

Parallel update applies (14) to d components simultaneously, and we may write the equivalent matrix form as $b^{(t+1)} = A_\gamma b^{(t)}$ for $A_\gamma = (I - \gamma S^{-1} V)$. On the contrary, sequential update applies (14) to one component at each iteration, and we may directly study the decrease in ELBO as it is quadratic (c.f. online supplementary material, Appendix D.6 for more details). It is worth noting that the parallel and the randomized sequential schemes using the preceding updating rule respectively coincide with the Jacobi and the Gauss–Seidel methods (with partial step size) for solving the system of d linear equations $Vx = 0$. To analyse the convergence of the algorithm, let us study the evolution of the objective functional $D(p^{(t)} \,\|\, \phi^*) = \frac{n}{2}[b^{(t)}]^T S b^{(t)}$, whose convergence implies the convergence of the algorithm. For the Gaussian posterior, we have $D^{(t)} = D(p^{(t)} \,\|\, \pi_n) - D(\phi^* \,\|\, \pi_n) = \frac{n}{2}[b^{(t)}]^T V b^{(t)}$, which is equivalent to $D(p^{(t)} \,\|\, \phi^*)$ up to some multiplicative constant. We summarize the result in the following lemma, whose proof is deferred to online supplementary material, Appendix D.6.

Lemma 3 (Gaussian posterior without latent variables). Suppose the true posterior is $N(\widehat{\theta}, (nV)^{-1})$. If the step size γ is chosen so that the α defined below belongs to $(0, 1)$, then $p^{(t)}$ converges exponentially fast to ϕ^*:

$$\mathbb{E}[D(p^{(t)} \,\|\, \phi^*)] \leq C D^{(t)} \leq C\alpha^t D^{(0)}, \quad t \geq 1,$$

for some constant $\alpha \in (0, 1)$ depending on the updating scheme. In particular, we have $\alpha = 1 - c_s\gamma(2 - \gamma)/d$ for the randomized sequential update; and $\alpha = c_p(\gamma)$ for the parallel update, where

$$c_s = \max_{\|b\|=1} \frac{b^T V S^{-1} V b}{b^T V b} \quad \text{and} \quad c_p(\gamma) = \max_{\|b\|=1} \frac{b^T (I - \gamma S^{-1} V) V (I - \gamma S^{-1} V) b}{b^T V b}. \tag{15}$$

Without parallel computing, the computational complexity of d iterations of the sequential scheme is comparable to that of the parallel scheme. Therefore, to fairly compare the overall computational complexities, we may define one iteration in the sequential scheme as d updates, whose effective contraction rate is then $(1 - c_s\gamma(2 - \gamma)/d)^d$, which is roughly $e^{-c_s\gamma(2-\gamma)}$ for a large d. Some detailed comparison of the contraction rates is deferred to the end of this subsection.

Remark 3 (Necessity of partial step size). From Lemma 3, we can see that the randomized sequential update always converges exponentially fast with full step size, i.e. $\alpha < 1$ for $\gamma = 1$. In fact, as $\alpha = 1 - c_s\gamma(2 - \gamma)/d$ in randomized sequential update

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

and $c_s > 0$ due to non-singularity of V, we may prefer taking $\gamma = 1$ to obtain the fastest convergence rate. In contrast, the parallel update may require a partial step size to guarantee the convergence. This is not an artefact of our proof technique—consider the following counterexample with $\gamma = 1$: we take $d = 3$, $V = \frac{1}{3}\boldsymbol{I}_3 + \frac{2}{3}\boldsymbol{1}_3 \cdot \boldsymbol{1}_3$ where $\boldsymbol{I}_3$ is the identity matrix and $\boldsymbol{1}_3$ the all-one vector. If initialized at $b^{(0)} = (1, 1, 1)$, the parallel scheme leads to $b^{(t)} = (-\frac{4}{3})^t b^{(0)}$, and therefore both $b^{(t)}$ and the ELBO diverges.

General case I: non-Gaussian posterior without latent variables. Recall that the CAVI output at tth iteration for a general posterior π_n is denoted as $q^{(t)}$, and the associated CAVI output for its large-sample normal approximation $N(\widehat{\theta}^{\text{mle}}_{\mathcal{M}}, [nV(\theta^*_{\mathcal{M}})]^{-1})$ (density denoted as ϕ_n) is $p^{(t)}$. In addition, the MF approximation to ϕ_n is $N(\widehat{\theta}, (nS)^{-1})$, whose density is denoted as ϕ^*; and $\widehat{q}$ is the MF approximation to π_n. Recall that $D^{(t)} = D(q^{(t)} \parallel \pi_n) - D(\widehat{q} \parallel \pi_n) = L(\widehat{q}) - L(q^{(t)})$ denotes the regret of $q^{(t)}$ relative to the theoretical ELBO value achieved by $\widehat{q}$. We prove the following theorem, describing the convergence of $D^{(t)}$ and $q^{(t)}$, by applying perturbation analysis for generalizing Lemma 3 to a general posterior.

Theorem 4 (General posteriors without latent variables). Under Assumptions 1 and 2, there exist positive constant C, such that if the support of initial distribution is in some small neighbourhood around θ^*, or $supp(q^{(0)}) \subseteq B_{\theta^*}(\delta) = \{\theta : \|\theta - \theta^*\| \leq \delta\}$ for some sufficiently small constant $\delta > 0$, then we have for any $1 \leq K = O(\sqrt{n})$ it holds with probability at least $1 - CK^{-2}$ that

$$\mathbb{E}[D^{(t)}] \leq C\alpha^t D^{(0)} + \frac{CK^3(\log n)^3}{\sqrt{n}}, \quad t \geq 1, \tag{16}$$

where the contraction rate α is given in Lemma 3. Moreover, under the same high-probability event, there exists some constant C_0 depending on the initialization $q^{(0)}$ such that

$$\mathbb{E}[H(q^{(t)}, \widehat{q})] \leq \alpha^{t/2} C_0 + C\sqrt{\frac{K^3(\log n)^3}{\sqrt{n}}}. \tag{17}$$

The proof of the theorem is quite technical and provided in online supplementary material, Appendix D.7, where the expression of C_0 is also included. As we can see, the general case inherits the same contraction rate α from the Gaussian case. However, the convergence is now only up to the $\mathcal{O}(\log^3(n)/\sqrt{n})$ statistical accuracy; and only a local geometric convergence can be proved. The local convergence is not an artefact of our proof, as the numerical example in Section 4.4 shows that the fast geometric convergence indeed only occurs after the iterate enters the contraction basin as a local neighbourhood around θ^*. We note that the $\log^3(n)/\sqrt{n}$ in our error bound is larger than the $\sqrt{\log(n)}/\sqrt{n}$ error bound proved by Alquier and Ridgway (2020) for the gradient-based method for MF inference. However, their result is only stated for the MF approximation to α-fractional posteriors with $\alpha < 1$, which requires fewer assumptions and tends to be technically less involved in proving. It would be interesting to explore if the same error bound also holds for the MF approximation to the usual posterior without tempering.

In view of Theorem 1, it suffices to run $O(d\log(nd))$ iterations to guarantee model selection consistency by using the approximated ELBO value produced from the CAVI algorithm. The first inequality in the theorem states that $D(q^{(t)} \parallel \pi_n)$ converges to $D(\widehat{q} \parallel \pi_n)$ exponentially fast up to a statistical error term of order $\mathcal{O}(\log^c n/\sqrt{n})$. However, the convergence of $D(q^{(t)} \parallel \pi_n)$ to $D(\widehat{q} \parallel \pi_n)$ (which is roughly equal to the constant $C^*(\mathcal{M})$, up to the same statistical error) does not imply the convergence of $q^{(t)}$ to $\widehat{q}$. Therefore, we also provide the second inequality that implies the convergence of $q^{(t)}$ towards $\widehat{q}$, where the loss function used is the Hellinger distance. We use the weaker Hellinger distance to characterize the convergence of $q^{(t)}$ by applying the triangle inequality because KL divergence is not a metric and hence triangle inequality is not applicable.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

General case II: non-Gaussian posterior with latent variables. When there are latent variables, we may apply similar arguments to analyse the convergence of the CAVI algorithm. The (parallel) updating formula in CAVI with latent variables, for example, becomes

$$q_{S_i}^{(t+1)}(s_i) \propto \exp\left\{\int_{\mathbb{R}^d} q_\theta^{(t)}(\theta)\log p\big(s_i \mid \theta, X_i\big)\,\mathrm{d}\theta\right\}, \quad i = 1, \ldots, n,$$

$$q_{\theta_j}^{(t+1)}(\theta_j) \propto \exp\left\{\int_{\mathbb{R}^{d-1}}\left(\int_{S^n} q_{S^n}^{(t+1)}(s^n)\log p(\theta \mid X^n, s^n)\,\mathrm{d}s^n\right) q_{\theta_{-j}}^{(t)}(\theta_{-j})\,\mathrm{d}\theta_{-j}\right\}, \quad j = 1, \ldots, d.$$

We will follow a similar analysis based on Taylor expansions on θ, first analysing $q_{S_i}^{(t+1)}$ by utilizing the concentration property of $q_\theta^{(t)}$ and then plugging it into $q_{\theta_j}^{(t+1)}$. This perturbation analysis helps us identify the leading terms in the CAVI updates, resulting in a key updating formula online supplementary material, (45) (in Appendix D.8) describing the 'noiseless' case dynamic, similar to the updating formula (14) for the Gaussian posterior without latent variables as obtained earlier. With this key updating formula, we can then proceed with the analysis as in the case without latent variables. First, we study the CAVI for the leading 'noiseless' Gaussian (or quadratic) case. We then employ perturbation analysis to analyse the CAVI for the original problem. The randomized sequential update can be interpreted and analysed in a similar way. This leads to the following theorem, whose proof is provided in online supplementary material, Appendix D.8. Recall that we used $V = V(\theta^*)$ to denote the observed data Fisher information, $V_s = V_s(\theta^*)$ for the missing data Fisher information, and $V_c = V + V_s$ for the complete data Fisher information. Additionally, S and S_c are the diagonal matrices corresponding to the diagonal parts of V and V_c, respectively.

Theorem 5 (General posteriors with latent variables). Under Assumptions 1, 2, and 3, the conclusions of Theorem 4 remain true, albeit with different contraction rates. In particular, we have $\alpha = 1 - c_s(\gamma)\gamma/d$ for the randomized sequential update; and $\alpha = c_p(\gamma)$ for the parallel update, where

$$c_s(\gamma) = \max_{\|b\|=1}\frac{b^T V S_c^{-1}(2S_c - \gamma S)S_c^{-1}Vb}{b^T V b} \quad \text{and}$$

$$c_p(\gamma) = \max_{\|b\|=1}\frac{b^T(I - \gamma S_c^{-1}V)V(I - \gamma S_c^{-1}V)b}{b^T V b}.$$

By comparing the contraction rates of the CAVI algorithm with latent variables (Theorem 5) to those without latent variables (Theorem 4), we can observe that the presence of latent variables generally slows down the convergence of CAVI. Specifically, this reduction in convergence rate is characterized by a factor related to the ratio between the observed data information and the complete data information, or the matrix operator norm of $S_c^{-1}S$, where diagonalization is due to the mean-field approximation on q_θ; see the sequential update case in online supplementary material, Appendix D.8 for a more concrete calculation. Our convergence result for the CAVI algorithm is consistent with existing findings on the algorithmic contraction rate of the EM algorithm (Dempster et al., 1977)—the convergence slows down as the missing data information V_s increases, making the latent variables harder to learn, which in turn affects the algorithmic convergence of the estimation of θ.

We end this section with a discussion on the efficiency comparison of the two updating schemes using an illustrating example without latent variables. We define an epoch for the sequential scheme to be d iterations so that each coordinate is updated once (on average for the randomized sequential update), and an epoch for the parallel scheme as one iteration. We compare the (averaged) decrease in ELBO for the two methods after one epoch. We consider $V = (1 - a)I_d + a1_d \cdot 1_d^T$. For the sequential update, $\mathbb{E}[D^{(t)}]$ has a relative decrease by at least $1 - (1 - \frac{\gamma(2-\gamma)(1-a)}{d})^d$, which will be $1 - (1 - (1 - a)/d)^d$ at $\gamma = 1$ after one epoch. For the parallel update, we need $\gamma((d - 1)a + 1) < 2$ to avoid divergence. Taking $\gamma = \frac{2}{(d-2)a+2}$ leads to a largest relative decrease by $1 - (1 - \frac{2-2a}{da-2a+2})^2$. When a is close to 1, we can see that the decrease from the parallel

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

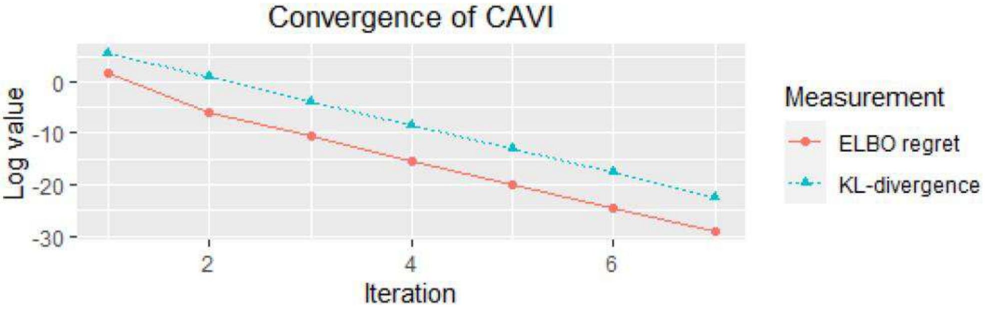

Figure 2. Convergence of CAVI for location-scale normal model.

update tends to be $4/d$ of that from the sequential update after one epoch. As a tends to zero, the relative decrease from the parallel update is about $1-(da/2)^2 \approx 1$, which is roughly 1.58 times larger (better) than the $1-e^{-1}$ decrease from the sequential update when d is large. Therefore, efficiency-wise, the parallel update is generally preferred in low-dependence situations (da small); while the sequential update is preferred in high-dependence and high-dimensional situations. Implementation-wise, the parallel update may benefit from parallel computing; the sequential update always converges while the parallel update may need fine tuning the step size.

4 Numerical study

This section includes numerical experiments on assessing the algorithmic convergence and the performance of using ELBO for model selection. Detailed descriptions about some of the considered examples are provided in online supplementary material, Appendix A.

4.1 Location-scale normal model

We consider the location-scale normal model $N(\mu, \sigma^2)$ with parameter $\theta=(\mu, \sigma^2)$ as described in online supplementary material, Appendix A. We generate sample from $N(100, 100^2)$ with sample size $n=10$. We take the prior as $\mu \sim N(0, 100^2)$ and $\sigma^2 \sim \mathrm{IG}(1/100, 1/100)$. To assess the convergence of the CAVI algorithm, we plot the logarithms of the ELBO regret $L(\widehat{q})-L(q^{(t)})$ and the KL divergence $D(q^{(t)} \,\|\, \widehat{q})$ between $q^{(t)}$ and $\widehat{q}$ respectively versus iteration count t. Note that the posterior has asymptotically independent components in this setting, so the two updating regimes lead to similar result, and we present sequential update for illustration. Here the optimal variational distribution $\widehat{q}$ is approximately obtained from the CAVI output after sufficiently many iterations. The results in Figure 2 show that both measurements are nearly linearly decreasing in the log-scale, implying the geometric convergence. The two curves are nearly parallel, which implies the same convergence rate. Note that the Fisher information V (equal to V_c without latent) in this example is diagonal. Therefore, the leading constant term of the contraction rate α from Theorem 4 vanishes and the remainder $\mathcal{O}(1/\sqrt{n})$ term characterizes the contraction rate.

In terms of the approximation accuracy of ELBO and BIC, the diagonal information matrix implies that the limiting constant gap $C^*(\mathcal{M})$ between evidence and ELBO equals 0, and they only differ by a statistical error term. On the other hand, the constant gap $C^*_{\mathrm{BIC}}(M)$ between evidence (or ELBO) and $-\mathrm{BIC}/2$ is generally nonzero. We can also see this difference via a numerical comparison, by plotting the relative approximation errors of ELBO and BIC under different sample sizes, defined as $\dfrac{ELBO-\log p(X\mid\mathcal{M})}{|\log p(X\mid\mathcal{M})|}$ and $\dfrac{\text{-BIC/2}-\log p(X\mid\mathcal{M})}{|\log p(X\mid\mathcal{M})|}$ respectively. The evidence is calculated based on 10^6 Monte Carlo replicates. The result is included in the left panel of Figure 3. To assess the impact of prior information on the approximation accuracy, we run an extra simulation study by fixing the sample size n to be 10, and taking the prior as $\mu \sim N(0, s^2)$ and $\sigma^2 \sim \mathrm{IG}(s^{-1}, s^{-1})$ with varying s over the grid $\exp(m)$ for $m=0, 0.2, \ldots, 5$. We report the result in the right panel of Figure 3. As expected, smaller s leads to larger approximation error of BIC, as BIC completely neglects the prior contribution in the evidence.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

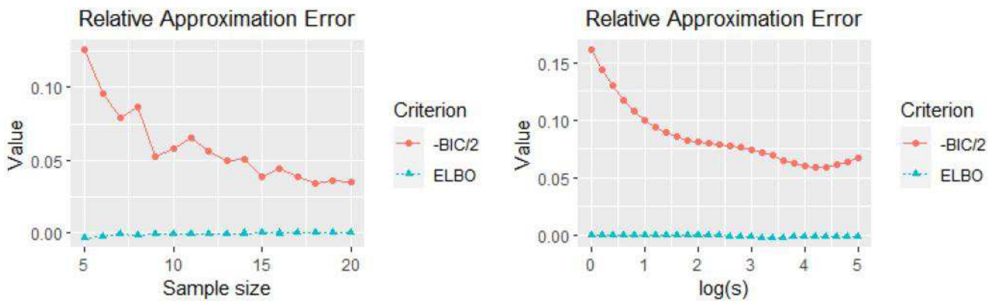

Figure 3. Relative approximation error of ELBO and -BIC/2 for normal parameters, with different sample sizes or different prior parameters.

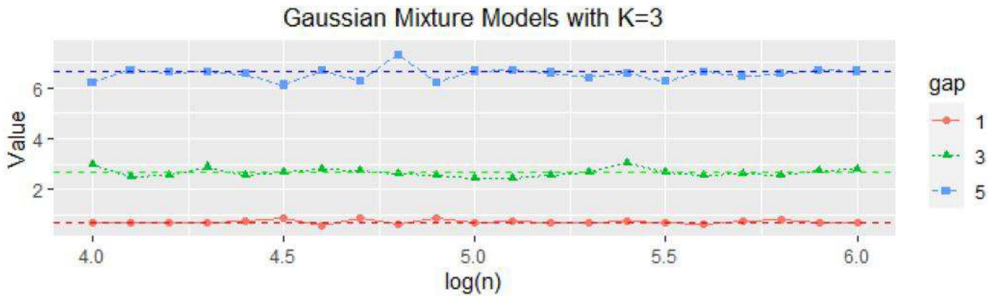

Figure 4. Comparison between simulated $-\text{BIC}/2 - \text{ELBO}$ and theoretical values $\widetilde{C}^*(\mathcal{M}_0)$ over three different gaps Δ. The dashed horizontal lines refer to theoretical values.

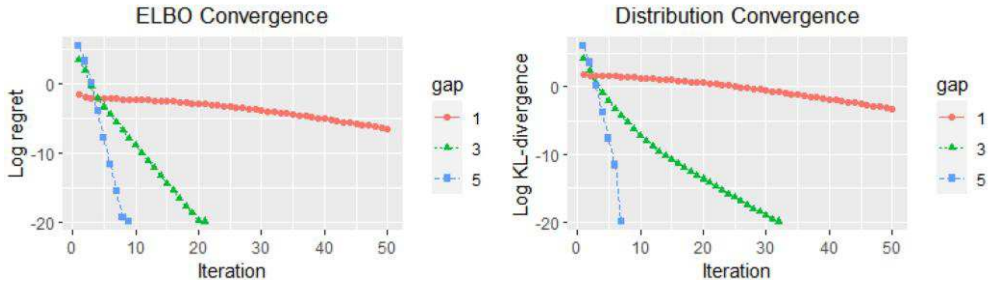

Figure 5. Convergence of CAVI in GMM with gap $\Delta = \{1, 3, 5\}$.

4.2 Gaussian mixture model

We consider the GMM with $K = 3$ clusters centred at $\Delta \cdot (-1, 0, 1)$ (a detailed description is provided in online supplementary material, Appendix A). We take the gap (or cluster separation) $\Delta \in \{1, 3, 5\}$, and the prior parameter $\sigma = 2$ in the i.i.d. $N(0, \sigma^2)$ prior for cluster centres (we choose relatively smaller σ so that $\widetilde{C}^*(\mathcal{M}_0)$ is more distinguishable in the plot for different Δ). We take the sample size n over the grid $\lfloor \exp(m) \rfloor$ for $m = 4, 4.1, \ldots, 7.9, 8$, and calculate $-\text{BIC}/2 - \text{ELBO}$ in comparison with theoretical values $\widetilde{C}^*(\mathcal{M}_0) = \Delta^2/\sigma^2 + K(\log \sigma^2 - \log K)/2$. We first focus on the GMM with a correctly specified number of clusters $K = 3$. We can see the simulated values are fairly close to the theoretical limit $\widetilde{C}^*(\mathcal{M}_0)$ from Figure 4.

We also assess the convergence of the sequential CAVI algorithm by plotting the logarithm of the ELBO regret and $D(q^{(t)} \,\|\, \widehat{q})$ under $\Delta \in \{1, 3, 5\}$, $n = 100$ and $\sigma = 10$. The numerical results from Figure 5 show the geometric convergence of both measurements as they are nearly linearly decreasing in the log-scale, which is consistent to the prediction from Theorem 4. The slopes of

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

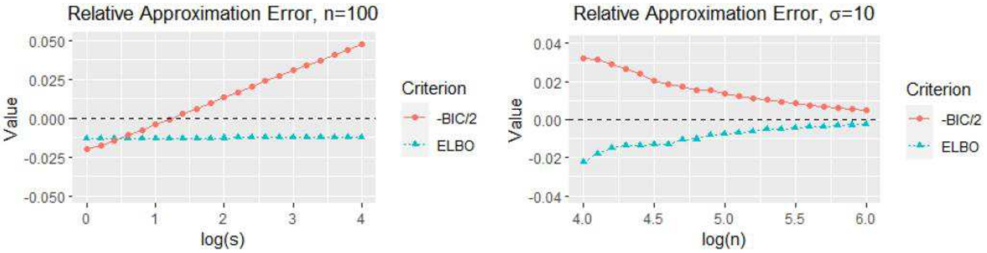

Figure 6. Approximation errors of ELBO and BIC in GMM under $\Delta = 1$.

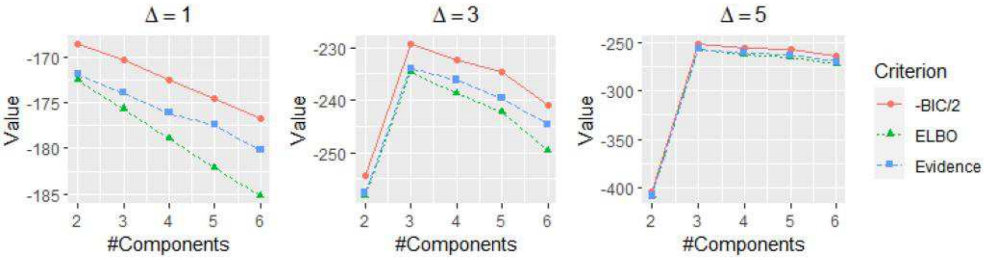

Figure 7. ELBO, BIC, and evidence values versus number of components in GMM.

two measurements are also close under the same Δ, implying the same convergence rate. Moreover, since a large gap Δ leads to a lower posterior dependence, our Theorem 4 predicts a smaller contraction rate α (faster convergence), which is also consistent with the numerical results. Next, we compare the use of BIC and ELBO for evidence approximation. We consider two settings: (1) prior standard deviation $\sigma = s$ for $s = \exp(m)$ with $m = 0, 0.2, \ldots, 4$ under fixed sample size $n = 100$ and (2) sample size $n = \lfloor \exp(m) \rfloor$ with $m = 4, 4.1, \ldots, 6$ with fixed $\sigma = 10$. We take the gap $\Delta = 1$. As we can see from Figure 6, the approximation error of BIC is quite sensitive to the prior choice; while the approximation of ELBO is much more accurate and stable. However, both approximations become accurate as we increase the sample size.

Note that GMM has degenerate Fisher information when the number of clusters K is misspecified, which leads to a slower $n^{-1/4}$ rate of convergence when K is overspecified (Chen, 1995). With known mixing weights, the overfitted models are still degenerate. For example, the two-component GMM with equal weights is degenerate when true model is a single Gaussian (Dwivedi et al., 2020). Although our theory no longer applies in such a situation, it is still interesting to compare the model selection performance based on BIC, ELBO, and the true evidence. For simplicity, we assume the underlying mixing weights to be known and fixed in our simulation setting. We set $\Delta \in \{1, 3, 5\}$, sample size $n = 100$ and prior with $\sigma = 10$. The results are summarized in Figure 7. As we can see, the three model selection criteria result in the same selected models across all settings. When the gap $\Delta = 1$ is relatively small, all criteria select the smaller model with two components. As we increase the gap to $\Delta \in \{3, 5\}$, all criteria select the true model. In terms of approximation accuracy for the evidence, we can see that ELBO tends to be more accurate than BIC, particularly for smaller models under stronger signal-to-noise ratios where the Fisher information is nonsingular near the misspecified MLE.

4.3 Generalized linear model

In this section, we consider variable selection in probit regression as a representative instance of GLM, whose detailed description is deferred to online supplementary material, Appendix A. The simulation setting is as follows. We generate p-dim feature vectors X_i i.i.d. from $N(0, \Sigma)$. The covariance matrix $\Sigma(r) = (\sigma_{ij}) \in \mathbb{R}^{p\times p}$ is of AR(1) structure so that $\sigma_{ij} = r^{|i-j|}$. We take $(n, d_{\mathcal{M}_0}) = (1{,}000, 10)$, and the p-dim (regression) coefficient $\boldsymbol{\beta}$ has the five nonzero terms:

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

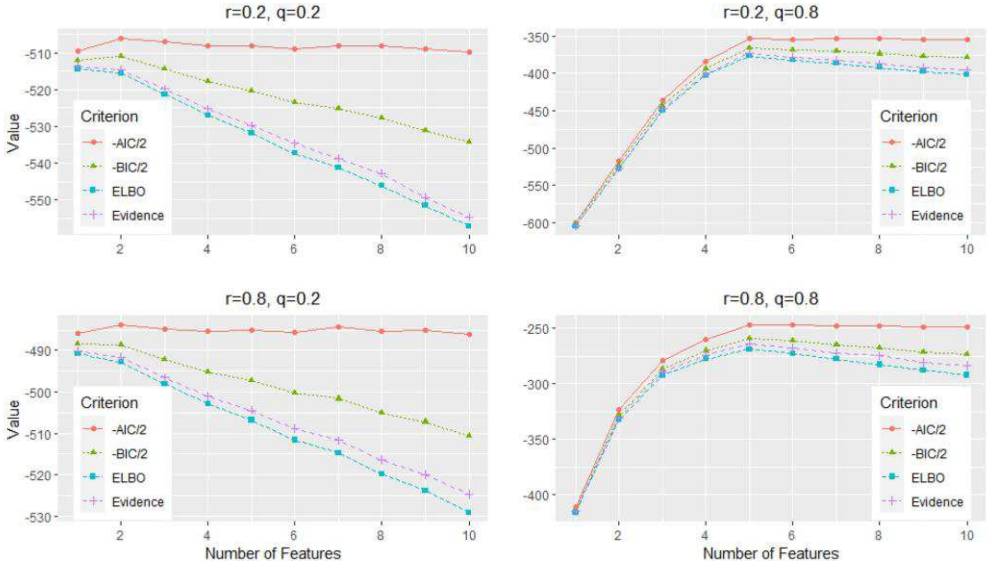

Figure 8. ELBO, AIC, BIC, and evidence values versus number of variables in probit regression.

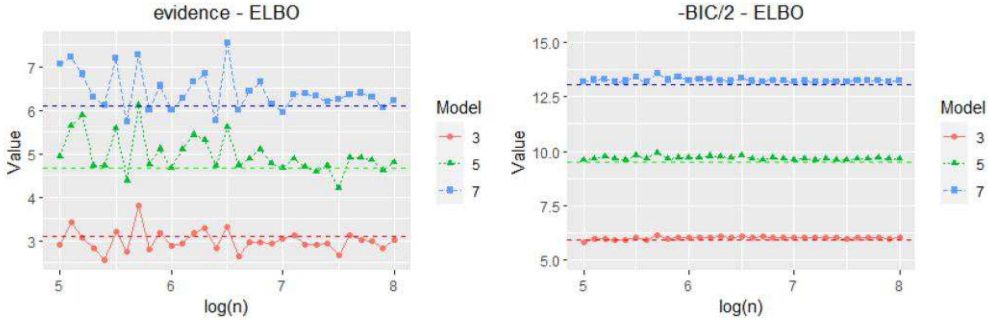

Figure 9. Comparison between simulated and theoretical values for $C^*(\mathcal{M})$ and $\widetilde{C}^*(\mathcal{M})$ under the models with $\{3, 5, 7\}$ included variables. The dashed horizontal lines refer to theoretical values.

$\beta_j = q^{j-1} \cdot 1(j \le 5)$, $j = 1, \ldots, 10$. The prior of $\boldsymbol{\beta}$ is the centred multivariate normal distribution with covariance matrix $\Sigma_0 = \sigma^2 I_{d_{\mathcal{M}}}$ where $\sigma = 10$. To assess the variable selection performance, we compare ELBO with AIC, BIC, and the true evidence (calculated based on 10^4 Monte Carlo replicates). As we can see from Figure 8, when the signal is relatively strong ($q = 0.8$), all methods except AIC correctly select the true model; and AIC selects a larger model (with six features) when the dependence is strong ($r = 0.8$). Moreover, the AIC curve is quite flat and the discrepancy between AIC values under different models is small. Under the weak signal case ($q = 0.2$), although the proposed criterion and BIC fail to select the true model, they still select the same model as that based on evidence (i.e. the highest posterior model). Under all four settings, ELBO is much more accurate than BIC in terms of the approximation for the evidence.

In addition, to compare the numerical gaps with the theoretical constants $C^*(\mathcal{M})$ and $\widetilde{C}^*(\mathcal{M})$ (which together also characterize $C^*_{\mathrm{BIC}}(\mathcal{M})$) in Theorem 2 and Ew. (13), Figure 9 includes their simulated values for the models including 3, 5, and 7 features (recall the true model contains five features), respectively, with the sample size n over the grid of $\lfloor \exp(m) \rfloor$ for $m = 5, 5.1, \ldots, 7.9, 8$. In this study, we set $q = 0.8$ and $r = 0.8$. We estimate $\boldsymbol{\beta}^*_{\mathcal{M}}$ for the underfitted model (with three features) based on 10^7 samples, so that the estimated $\widehat{\boldsymbol{\beta}}_{\mathcal{M}}$ can be treated as $\boldsymbol{\beta}^*_{\mathcal{M}}$. As we can see from Figure 9, the simulated values are fairly close to the theoretical ones, which justifies our theory that

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

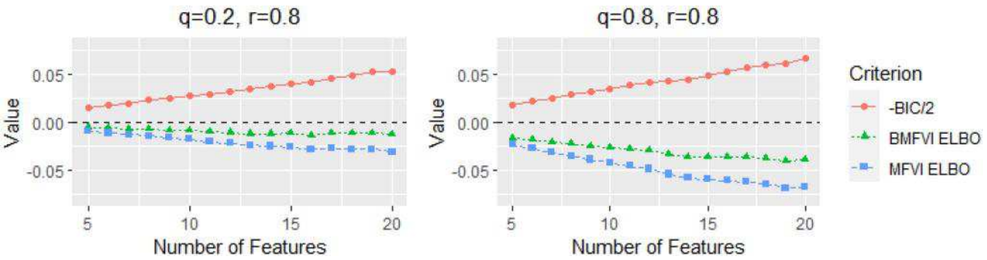

Figure 10. Relative approximation errors of ELBO and -BIC/2 under weak ($q = 0.2$) and strong ($q = 0.8$) signal regimes.

evidence −ELBO and −BIC/2 − ELBO are equal to $C^*(\mathcal{M})$ and $\widetilde{C}^*(\mathcal{M})$ up to some higher-order terms. We note that the simulated values of $C^*(\mathcal{M})$ have higher fluctuations, which might be partially caused by the high variance of the Monte Carlo approximation to the evidence.

According to Theorem 2, the difference between -BIC/2 and evidence has a term proportional to the dimension $d_{\mathcal{M}}$, which is consistent with the trend of the gap observed from Figure 8. On the other hand, as we used block MF (BMFVI) for probit regression, the approximation error of ELBO only comes from the difference between V_c and V characterized by $\frac{1}{2}\log\frac{\det(V_c)}{\det(V)}$, which grows linearly in $d_{\mathcal{M}}$ with slope independent of σ_0. To formally demonstrate the impact of $d_{\mathcal{M}}$ on approximation the evidence, we plot the approximation errors of BIC and ELBO for $d_{\mathcal{M}} = 5, \ldots, 20$ in Figure 10. In the plot, we also include the results for the fully factorized MF (MFVI), whose limiting approximation error $\frac{1}{2}\log\frac{\det(\operatorname{diag}(V_c))}{\det(V)}$ tends to be more sensitive to $d_{\mathcal{M}}$ than BMFVI, as expected. We can see that ELBO with BMFVI has overall best approximation accuracy, while the approximation error of BIC is the largest due to its strong dimension dependence. On the other hand, the ELBO from MFVI has larger approximation error than that from BMFVI, due to the additional independence imposed among parameter components. Therefore, we should take advantage of the block mean-field structure and implement BMFVI in practice whenever computationally tractable.

To assess the prediction performance of various criteria in the presence of weak signals (i.e. selection consistency may not be achievable) in the context of Theorem 3, we consider a more challenging setting where $d_{\mathcal{M}_0} = 100$, $\beta_j = 0.8^j$ for $1 \le j \le d_{\mathcal{M}_0}$, and the same covariance matrix $\Sigma(r)$ with $r \in \{0.2, 0.8\}$ for the random feature vectors. Although now the true model includes all 100 important features, signals decay geometrically fast; as a result, both BIC and ELBO would select much smaller models. Although model selection consistency does not hold under this setting, we use the numerical study to compare the prediction performance of model selection based on AIC, BIC, and ELBO. In this example, we include AIC as the benchmark corresponding to the theoretically minimax-optimal criterion for prediction (Hirotugu, 1974). We generate 10^4 data, and randomly select $n = 200, 500, 1{,}000$ of them as training data, and the remaining as test data. Then we select the model using the training data with AIC, BIC, and ELBO, and evaluate their prediction performance on the test data. We compare the classification errors and logistic losses of ELBO and BIC evaluated on the test data. Note that the logistic loss is defined as $-n_{\text{test}}^{-1}\sum_{i=1}^{n_{\text{test}}}[y_i\log\widehat{p}_i + (1-y_i)\log(1-\widehat{p}_i)]$, where $\widehat{p}_i$ is the fitted probability on the ith test data with the selected model. The logistic loss is the log-likelihood function evaluated on the test data, and approximates the population-level KL divergence between the true and the fitted prediction distributions up to some fixed constant. We report the average classification errors along with their standard deviations, and the median of logistic losses from 100 Monte Carlo replicates. From Table 1, the prediction accuracy of ELBO is overall comparable to AIC, and outperforms BIC. However, AIC tends to include more variables than ELBO for slightly reducing its classification error; while due to the tendency of overfitting the data by including too many variables, AIC may incur slightly larger logistic loss on the test data. Overall, ELBO seems to be a better balance between AIC and BIC by selecting a parsimonious model with comparable prediction accuracy as AIC.

Lastly, we assess the algorithmic convergence of the two CAVI algorithms for implementing MFVI. We take the true model with $n = 100$ samples and $d_{\mathcal{M}_0} = 10$ features with $\boldsymbol{\beta}^* = 0.1 \cdot \mathbf{1}_{10}$,

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Table 1. Prediction comparison among ELBO, AIC, and BIC in probit regression

r	n	Classification error in %			Logistic loss			Average model size		
		ELBO	AIC	BIC	ELBO	AIC	BIC	ELBO	AIC	BIC
0.2	200	21.5 (1.6)	20.7 (1.6)	24.1 (2.7)	0.421	0.430	0.451	5.2 (0.9)	7.8 (1.6)	3.6 (0.9)
	500	18.8 (0.7)	18.5 (0.8)	20.0 (1.1)	0.363	0.366	0.381	7.4 (0.9)	9.8 (1.5)	5.7 (0.8)
	1,000	18.0 (0.5)	17.8 (0.5)	18.7 (0.6)	0.346	0.346	0.356	8.7 (0.9)	11.2 (1.4)	7.2 (0.8)
0.8	200	13.3 (1.0)	13.0 (1.0)	14.6 (1.5)	0.276	0.282	0.289	3.8 (0.9)	4.8 (1.2)	2.7 (0.6)
	500	11.4 (0.6)	11.2 (0.7)	12.5 (0.9)	0.229	0.227	0.248	5.5 (0.9)	6.9 (1.2)	4.0 (0.6)
	1,000	10.7 (0.5)	10.5 (0.5)	11.3 (0.5)	0.215	0.211	0.225	7.0 (0.9)	8.6 (1.0)	5.4 (0.8)

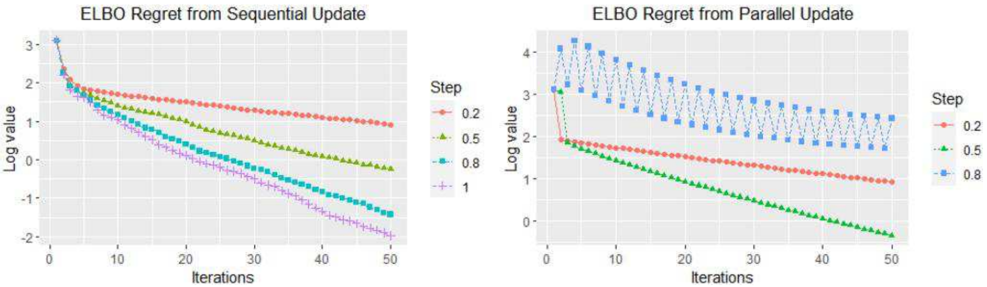

Figure 11. ELBO convergence in probit regression under sequential and parallel updating schemes.

and feature covariance matrix $\Sigma = (\sigma_{ij})_{i,j=1}^{p}$ with $\sigma_{ij} = 0.9 + 0.1 \times 1(i = j)$. The prior standard deviation is $\sigma_0 = 1$ and $\boldsymbol{\beta}$ initialized at $\boldsymbol{\beta}^{(0)} = 0_{10}$. Figure 11 reports the logarithm of ELBO regret with different step sizes from parallel and randomized sequential updates. As we can see, the sequential update is always convergent and has faster convergence rate with a larger step size; moreover, the log ELBO regret has a linear decay, imply the geometric convergence of ELBO. On the contrary, the parallel update only converges under smaller step sizes—the ELBO converges extremely slow due to oscillations or even diverges to ∞ under larger or full step size (hence not presented), and $\gamma = 0.8$ tends to be the threshold under which CAVI converges fast without oscillations. Consistent with our theory, the parallel CAVI has geometric convergence for small step sizes $\gamma = 0.2$ and 0.5. In addition, for both updating schemes, the slope (equal to $\log \alpha$) of the log regret is roughly proportional to the step size in absolute value (except for $\gamma = 0.8$ in the parallel update).

4.4 Community detection

In this example, we consider the model selection problem of selecting the number of communities in the community detection described in online supplementary material, Appendix A. We set the true underlying data generating SBM to have five communities. The five communities have equal number of members. The connectivity matrix $\boldsymbol{B}$ is generated as $\boldsymbol{B}_{ab} \overset{iid}{\sim} \mathrm{Unif}(0, 0.4)$ if $i \neq j$ and $\boldsymbol{B}_{aa} = 0.6$. For candidate models, we consider $K = 2, \ldots, 10$ communities. We take the priors on $\boldsymbol{B}_{ab}$ as Beta(1,1) and on Z_i as Categorical$(1/K, \ldots, 1/K)$. We calculate the ELBO using the parallel update in CAVI with full step size, since we find that the algorithm always converges. In terms of BIC, since we cannot find the exact MLE, we approximate MLE by the variational mean from $\widehat{q}_\theta$ and obtain the BIC correspondingly.

For SBM, the Fisher information is infeasible as is for MLE. However, we can visually find $\widetilde{C}^*(\mathcal{M})$, i.e. the gap between -BIC/2 and ELBO via the numerical differences. Note that it is no longer a 'constant' gap since we have growing number of parameters in SBM. We have included the ELBO and BIC values obtained with $n = 100$ and 1,000 nodes in Figure 12. Both criteria select the true model under both settings, but the gap between two criteria is comparable to the gap between

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

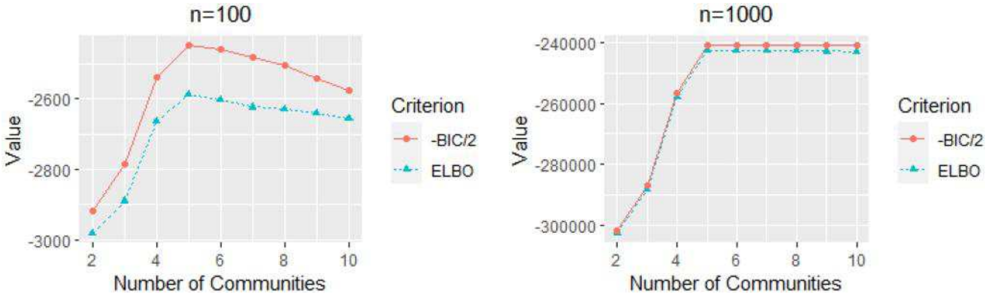

Figure 12. ELBO and BIC from parallel CAVI in SBM.

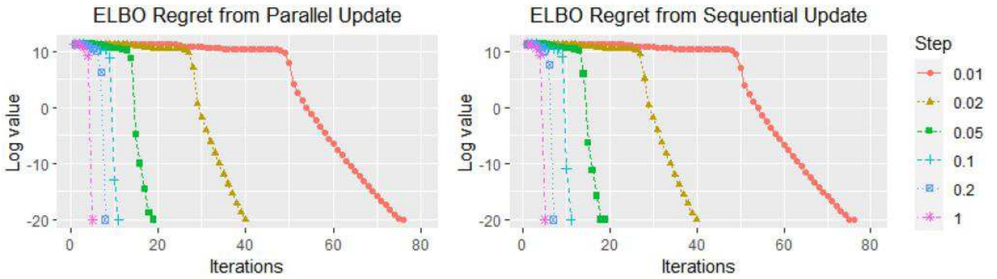

Figure 13. ELBO and BIC from parallel CAVI in SBM with $n = 1{,}000$.

different models, particularly for the overfitted model. This is due to the growing number of parameters.

We plot in Figure 13 the regret of ELBO versus iteration counts from both parallel and sequential CAVI algorithms under the true model, with different step sizes. We can see that two updating regimes have overall similar performance. The ELBO decays slowly in the first several iterations, and converges exponentially fast afterwards. The convergence speed (slope) is roughly proportional to the step size and the two updating regimes have almost identical speed under small step sizes. This is consistent with our findings in Theorem 4 that CAVI has geometric convergence given a good initialization, as the ELBO in SBM starts to converge exponentially fast after entering the local contraction basin around the true parameter.

4.5 Real data application

In this section, we apply and compare the three model selection criteria, namely ELBO, AIC, and BIC, in four classification datasets: Adult, Taiwan Company Bankruptcy (TRB), Musk, and Car Insurance Claims (CIC). The first three are publicly available on the UCI machine-learning repository, and the CIC data are available at Kaggle.

The original Adult dataset contains 14 features, including 6 continuous and 8 categorical ones. Instead, we use the preprocessed version on the LibSVM repository containing 123 binary features and 32,561 samples. The response variable is a binary one encoding whether the individual has a salary ≥50 K. The TRB dataset contains 95 features and 6,819 samples. We remove one constant feature and use the remaining to predict whether the company will go bankruptcy, based on the business regulations of the Taiwan Stock Exchange. Since the data are highly unbalanced as only 220 data points have response one (referring to bankruptcy), we repeat each of them for additional 10 times so the derived dataset contains 9,019 samples. This is equivalent to a weighted probit regression. For the Musk dataset about molecules, we use its second version that contains 6,598 samples and 166 features. The goal is to predict whether the molecules are musks or not. The original CIC dataset contains 10,000 samples and 17 features. We remove the samples

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Table 2. Comparison between ELBO, AIC, and BIC in real data applications

Data	Classification error in %			Logistic loss			Average model size		
	ELBO	AIC	BIC	ELBO	AIC	BIC	ELBO	AIC	BIC
Adult	18.6 (1.3)	18.2 (0.8)	20.0 (1.9)	0.369	0.366	0.384	7.0 (1.3)	11.0 (1.8)	3.7 (0.8)
TRB	14.9 (0.7)	14.4 (0.7)	15.5 (0.8)	0.328	0.353	0.317	5.8 (1.7)	8.7 (1.9)	3.4 (0.7)
Musk	10.2 (1.1)	8.6 (0.7)	12.2 (1.3)	0.265	0.243	0.290	10.6 (3.1)	17.7 (3.5)	5.5 (1.3)
CIC	17.0 (0.8)	16.9 (0.6)	17.9 (1.6)	0.333	0.330	0.349	10.2 (1.2)	11.0 (1.1)	8.0 (1.0)

containing missing values and apply one-hot encoding to the categorical features, and the preprocessed data contain 8,149 samples and 26 features.

We compare the prediction performance in terms of classification error, logistic loss, and selected model size, similar to the numerical study in Section 4.3. For each dataset, we randomly select 1,000 samples as the training data, which is used to conduct model selection with ELBO, AIC, and BIC. The selected models are fitted using the same 1,000 training data and evaluated on the remaining data as the test data. This procedure is repeated 100 times, and we report the average classification errors and median logistic losses, and also the average sizes of the models selected by different criteria. As we can see from Table 2, AIC has the overall best performance in terms of prediction as expected, which is consistent with the optimality of AIC in prediction accuracy (Yang, 2005). On the other hand, the performance of the proposed ELBO criterion tends to be in the between of AIC and BIC, with slightly worse prediction performance than AIC but better than BIC. The model size selected by ELBO also lies between AIC and BIC.

5 Conclusion

In summary, we proved a non-asymptotic BvM-type theorem for the MF variational distribution, stating that the variational distribution concentrates to a normal distribution centred at MLE with diagonal covariance matrix. Motivated by this normal approximation, we proposed a model selection criterion based on ELBO, and studied the model selection consistency. We found that BIC, ELBO, and evidence tend to differ from each other by some constant gaps; and ELBO tends to provide a better approximation to evidence than BIC due to a better dimension dependence and the full incorporation of the prior information. To characterize prediction efficiency, we also proved an oracle inequality for the variational distribution from the model selected by ELBO. Lastly, we showed the geometric convergence of two commonly used CAVI algorithms for solving the optimization problem in MF approximation; and the analysis suggested that it suffices to run $O(d\log(nd))$ iterations of CAVI algorithm to approximate the ELBO in order to achieve model selection consistency.

Next, we discuss some possible future directions. First, note that we have assumed the nonsingularity of the Fisher information to develop the theory. It would be interesting to study the model selection based on ELBO for singular models, such as the over-specified Gaussian mixture model numerically examined in Section 4.2, and study the connection between ELBO and the so-called singular BIC as discussed by Drton et al. (2017) and Watanabe and Opper (2010). Although our current theory on the consistent estimation (up to a constant) of model evidence using ELBO applies only to underspecified GMM with a nonsingular Fisher information, our numerical results for GMM suggest that the ELBO is still capable of identifying the true model. On the theoretical front, for a singular model $\mathcal{M}$, it is proved (see, for instance, Drton et al., 2017; Watanabe & Opper, 2010) that the model evidence $\log p(X^n|\mathcal{M})$ can be approximated by the singular BIC, or $\mathrm{sBIC}(\mathcal{M}) = -\widehat{\ell}_n(\mathcal{M}) + \lambda(\mathcal{M})\log n - [m(\mathcal{M}) - 1]\log\log n$ up to a bounded constant. Here, $\lambda(\mathcal{M})$ is called the learning rate or real log-canonical threshold of model $\mathcal{M}$ which summarizes the effective model dimensionality, and $m(\mathcal{M})$ represents the multiplicity of $\lambda(\mathcal{M})$. For regular models with invertible Fisher information matrices, $\lambda(\mathcal{M})$ simplifies to $d_{\mathcal{M}}/2$ and $m(\mathcal{M}) = 1$. As such, sBIC reduces to half of the usual BIC defined in Eq. (6). Recent work by Bhattacharya et al. (2020) demonstrates that the ELBO from mean-field variational inference accurately

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

captures the leading $\lambda(\mathcal{M}) \log n$ term for any singular model in its canonical or normal-crossing configuration. However, while Hironaka's theorem in algebraic geometry guarantees the existence of an analytical map (or reparametrization map) for any singular model that can locally transform the posterior distribution into a normal-crossing form, it is typically challenging to explicitly construct such an analytical map. In fact, the determination of $\lambda(\mathcal{M})$ and $m(\mathcal{M})$ for a generic singular model $\mathcal{M}$ is still an unresolved issue. As such, investigating whether the ELBO from MF variational inference can recover the leading $\lambda(\mathcal{M}) \log n$ term for a general singular model not in its normal-crossing from is an interesting question to explore. Second, in more complicated models, the latent variables and sample points may not be one-on-one. For example, if we treat the assignment variables as latent variables instead of unknown parameters in mixed membership community models, then we have n latent variables and $\binom{n}{2}$ data. Our theories do not cover these settings and we also leave them to future study. Finally, it would be interesting to extend the current development to models with dependent latent variables, such as hidden Markov models or state space models.

Acknowledgments

We are grateful to two anonymous reviewers for their constructive comments, which have led to substantial improvements of this paper.

Conflict of interest: To the best of our knowledge, this paper does not have any conflict of interest.

Funding

Y.Y.'s research was supported in part by National Science Foundation DMS-2210717.

Data availability

The four datasets we used in this paper include Adult, Taiwan Company Bankruptcy (TRB), Musk, and Car Insurance Claims (CIC). All of them are publicly available, with the first three on the UCI machine learning repository, and CIC at Kaggle.

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series B*.

References

Alquier P., & Ridgway J. (2020). Concentration of tempered posteriors and of their variational approximations. *The Annals of Statistics*, *48*(3), 1475–1497. https://doi.org/10.1214/19-AOS1855

Alquier P., Ridgway J., & Chopin N. (2016). On the properties of variational approximations of Gibbs posteriors. *The Journal of Machine Learning Research*, *17*(236), 8374–8414.

Bhattacharya A., Pati D., & Plummer S. (2020). 'Evidence bounds in singular models: Probabilistic and variational perspectives', arXiv, arXiv:2008.04537, preprint: not peer reviewed.

Bhattacharya A., Pati D., & Yang Y. (2019). Bayesian fractional posteriors. *The Annals of Statistics*, *47*(1), 39–66. https://doi.org/10.1214/18-AOS1712

Bickel P., Choi D., Chang X., & Zhang H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, *41*(4), 1922–1943. https://doi.org/10.1214/13-AOS1124

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Blei D. M., Kucukelbir A., & McAuliffe J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, *112*(518), 859–877. https://doi.org/10.1080/01621459.2017.1285773

Chen J. (1995). Optimal rate of convergence for finite mixture models. *The Annals of Statistics*, *23*(1), 221–233. https://doi.org/10.1214/aos/1176324464

Chérief-Abdellatif B.-E. (2019). Consistency of ELBO maximization for model selection. In *Symposium on advances in approximate Bayesian inference* (pp. 11–31). PMLR.

Chérief-Abdellatif B.-E., & Alquier P. (2018). Consistency of variational Bayes inference for estimation and model selection in mixtures. *Electronic Journal of Statistics*, *12*(2), 2995–3035. https://doi.org/10.1214/18-EJS1475

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Dempster A. P., Laird N. M., & Rubin D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Drton M., Plummer M., Claeskens G., Rousseau J., Robert C. P., Imai T., Kuroki M., Friel N., McKeone J., Oates C., Pettitt A. N., McLachlan G. J., Nguyen H. D., Charnigo R., Longford N. T., Bhansali R. J., Cameron E., Pettitt A. N., Chopin N., …Kuriki S. (2017). A Bayesian information criterion for singular models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *79*(2), 323–380. https://doi.org/10.1111/rssb.12187

Dwivedi R., Ho N., Khamaru K., Wainwright M. J., Jordan M. I., & Yu B. (2020). Singularity, misspecification and the convergence rate of EM. *The Annals of Statistics*, *48*(6), 3161–3182. https://doi.org/10.1214/19-AOS1924

Gelfand A. E., & Smith A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(410), 398–409. https://doi.org/10.1080/01621459.1990.10476213

Ghosal S., Ghosh J. K., & Van Der Vaart A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics*, *28*(2), 500–531.https://doi.org/10.1214/aos/1016218228

Ghosal S., & Van Der Vaart A. (2007). Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, *35*(1), 192–223. https://doi.org/10.1214/009053606000001172

Hall P., Ormerod J. T., & Wand M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, *12*(1), 369–389.

Hall P., Pham T., Wand M. P., & Wang S. S. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *The Annals of Statistics*, *39*(5), 2502–2532. https://doi.org/10.1214/11-AOS908

Hammersley J. (2013). *Monte Carlo methods*. Berlin, Germany: Springer Science & Business Media.

Han W., & Yang Y. (2019). 'Statistical inference in mean-field variational Bayes', arXiv, arXiv:1911.01525, preprint: not peer reviewed.

Hastings W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications.

Hirotugu A. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Ho N., & Nguyen X. (2019). Singularity structures and impacts on parameter estimation in finite mixtures of distributions. *SIAM Journal on Mathematics of Data Science*, *1*(4), 730–758. https://doi.org/10.1137/18M122947X

Jain V., Koehler F., & Mossel E. (2018). The mean-field approximation: Information inequalities, algorithms, and complexity. In *Conference on learning theory* (pp. 1326–1347). PMLR.

Jordan M. I., Ghahramani Z., Jaakkola T. S., & Saul L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233. https://doi.org/10.1023/A:1007665907178

Kleijn B. J. K., & Van der Vaart A. W. (2012). The Bernstein-Von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6(none), 354–381. https://doi.org/10.1214/12-EJS675

Koehler F. (2019). Fast convergence of belief propagation to global optima: Beyond correlation decay. *Advances in Neural Information Processing Systems*, *32*

Mukherjee S. S., Sarkar P., Wang Y., & Yan B. (2018). Mean field for the stochastic blockmodel: Optimization landscape and convergence issues. *Advances in Neural Information Processing Systems*, *31*

Ohn I., & Lin L. (2021). 'Adaptive variational Bayes: Optimality, computation and applications', arXiv, arXiv:2109.03204, preprint: not peer reviewed.

Ormerod J. T., & Wand M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *Journal of Computational and Graphical Statistics*, *21*(1), 2–17. https://doi.org/10.1198/jcgs.2011.09118

Pati D., Bhattacharya A., & Yang Y. (2018). On statistical optimality of variational Bayes. In *International conference on artificial intelligence and statistics* (pp. 1579–1588). PMLR.

Plummer S., Pati D., & Bhattacharya A. (2020). Dynamics of coordinate ascent variational inference: A case study in 2D Ising models. *Entropy*, *22*(11), 1263. https://doi.org/10.3390/e22111263

Saeedi A., Kulkarni T. D., Mansinghka V. K., & Gershman S. J. (2017). Variational particle approximations. *The Journal of Machine Learning Research*, *18*(69), 2328–2356.

Sarkar P., Wang Y. R., & Mukherjee S. S. (2021). When random initializations help: A study of variational inference for community detection. *Journal of Machine Learning Research*, *22*(22), 1–46.

Schwarz G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Shen X., & Wasserman L. (2001). Rates of convergence of posterior distributions. *The Annals of Statistics*, *29*(3), 687–714. https://doi.org/10.1214/aos/1009210686

Stoica P., & Selen Y. (2004). Model-order selection: A review of information criterion rules. *IEEE Signal Processing Magazine*, *21*(4), 36–47. https://doi.org/10.1109/MSP.2004.1311138

Titsias M., & Lázaro-Gredilla M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International conference on machine learning* (pp. 1971–1979). PMLR.

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024

Titterington D., & Wang B. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, *1*(3), 625–650. https://doi.org/10.1214/06-BA121

Trefethen L. N., & Bau III D. (1997). *Numerical linear algebra* (Vol. 50). Philadelphia, PA: Siam.

Van der Vaart A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge, UK: Cambridge University Press.

Wang B., & Titterington D. M. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *International workshop on artificial intelligence and statistics* (pp. 373–380). PMLR.

Wang Y., & Blei D. (2019). Variational Bayes under model misspecification. *Advances in Neural Information Processing Systems*, *32*, 13357–13367.

Wang Y., & Blei D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association*, *114*(527), 1147–1161. https://doi.org/10.1080/01621459.2018.1473776

Wang Y., Chen J., Liu C., & Kang L. (2021). Particle-based energetic variational inference. *Statistics and Computing*, *31*(1), 1–17. https://doi.org//10.1007/s11222-021-10009-7

Watanabe S., & Opper M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, *11*(116), 3571–3594.

Westling T., & McCormick T. H. (2015). 'Establishing consistency and improving uncertainty estimates of variational inference through M-estimation', arXiv, arXiv:1510.08151, preprint: not peer reviewed, **1**.

Wright S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, *151*(1), 3–34. https://doi.org/10.1007/s10107-015-0892-3

Yang Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika*, *92*(4), 937–950. https://doi.org/10.1093/biomet/92.4.937

Yang Y., Pati D., & Bhattacharya A. (2020). α-variational inference with statistical guarantees. *The Annals of Statistics*, *48*(2), 886–905. https://doi.org/10.1214/19-AOS1827

Yang Y., Wainwright M. J., & Jordan M. I. (2016). On the computational complexity of high-dimensional Bayesian variable selection. *The Annals of Statistics*, *44*(6), 2497–2532. https://doi.org/10.1214/15-AOS1417

Yin M., Wang Y. R., & Sarkar P. (2020). A theoretical case study of structured variational inference for community detection. In *International conference on artificial intelligence and statistics* (pp. 3750–3761). PMLR.

Zhang A. Y., & Zhou H. H. (2020). Theoretical and computational guarantees of mean field variational inference for community detection. *The Annals of Statistics*, *48*(5), 2575–2598. https://doi.org/10.1214/19-AOS1898

Zhang F., & Gao C. (2020). Convergence rates of variational posterior distributions. *The Annals of Statistics*, *48*(4), 2180–2207. https://doi.org/10.1214/19-AOS1883

Downloaded from https://academic.oup.com/jrsssb/article/86/3/742/7642825 by University of Maryland - College Park user on 23 September 2024