Variational Bayes with Intractable Likelihood

Minh-Ngoc Tran, David J. Nott, Robert Kohn*

Abstract

Variational Bayes (VB) is rapidly becoming a popular tool for Bayesian inference in statistical modeling. However, the existing VB algorithms are restricted to cases where the likelihood is tractable, which precludes the use of VB in many interesting situations such as in state space models and in approximate Bayesian computation (ABC), where application of VB methods was previously impossible. This paper extends the scope of application of VB to cases where the likelihood is intractable, but can be estimated unbiasedly. The proposed VB method therefore makes it possible to carry out Bayesian inference in many statistical applications, including state space models and ABC. The method is generic in the sense that it can be applied to almost all statistical models without requiring too much model-based derivation, which is a drawback of many existing VB algorithms. We also show how the proposed method can be used to obtain highly accurate VB approximations of marginal posterior distributions.

Keywords. Approximate Bayesian computation, marginal likelihood, natural gradient, quasi-Monte Carlo, state space models, stochastic optimization.

*Minh-Ngoc Tran is Lecturer, The University of Sydney Business School, Sydney 2006 Australia (minh-ngoc.tran@sydney.edu.au). David J. Nott is Associate Professor, Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 (standj@nus.edu.sg). Robert Kohn is Professor, UNSW Business School, University of New South Wales, Sydney 2052 Australia (r.kohn@unsw.edu.au).

1 Introduction

Let y be the data and $\theta \in \Theta$ the vector of model parameters. We are interested in Bayesian inference about θ, based on the posterior distribution $\pi(\theta) = p(\theta|y) \propto p(\theta)p(y|\theta)$, with $p(\theta)$ the prior and $p(y|\theta)$ the likelihood function. In this paper, we are interested in Variational Bayes (VB), which is widely used as a computationally effective method for approximating the posterior distribution $\pi(\theta)$ (Attias, 1999; Bishop, 2006). VB approximates the posterior by a distribution $q(\theta)$ within some tractable class, such as an exponential family, chosen to minimize the Kullback-Leibler divergence between $q(\theta)$ and $\pi(\theta)$. Most of the VB algorithms in the literature require that the likelihood $p(y|\theta)$ can be computed analytically for any θ.

In many applications, however, the likelihood $p(y|\theta)$ is intractable in the sense that it is infeasible to compute $p(y|\theta)$ exactly at each value of θ, which makes it difficult to use VB for inference. For example, in state space models (Durbin and Koopman, 2001), which are widely used in economics, finance and engineering, the likelihood is a high dimensional integral over the state variables governed by a Markov process. Ghahramani and Hinton (2000) were the first to use VB for inference in state space models. However, they only consider the special case in which the time series is segmented into regimes with each regime assumed to follow a linear-Gaussian state space model. For general state space models, it is still a challenging problem to do inference with VB. Turner and Sahani (2011) discuss some of the difficulties in applying VB methods to time series models. Another example of a situation where implementing VB is difficult is approximate Bayesian computation (ABC) (Tavare et al., 1997; Marin et al., 2012; Peters et al., 2012). ABC methods provide a way of approximating the posterior $\pi(\theta)$ when the likelihood is difficult to compute but it is possible to simulate data from the model. We are not aware of any work that uses VB for inference in ABC, although a closely related technique called Expectation Propagation has been used (Barthelme and Chopin, 2014). This paper proposes a VB algorithm that approximates $\pi(\theta)$ when the likelihood is intractable. The only requirement is that the intractable likelihood can be estimated unbiasedly. The proposed algorithm therefore makes it possible to carry out variational Bayes inference in many statistical models with an intractable likelihood, where this was previously impossible.

In many models, by introducing a latent variable α, the joint density $p(y,\alpha|\theta)$ is tractable. This makes it much easier to work with the joint posterior $p(\theta,\alpha|y) \propto p(\theta)p(y,\alpha|\theta)$ rather than the marginal posterior of interest $\pi(\theta)$ itself. In this situation many VB algorithms in the literature approximate the joint posterior $p(\theta,\alpha|y)$ by a factorized distribution $q(\theta)q(\alpha)$, and then use $q(\theta)$ as an approximation to $\pi(\theta)$. The main drawback of this approach is that the (usually high) posterior dependence between θ and α is ignored, which might lead to a poor VB approximation (Neville et al., 2014). Our VB algorithm approximates $\pi(\theta)$ directly with the latent variable α integrated out and thus overcomes this drawback; see the example in Section 5.1.

Section 2 presents our approach, which we call Variational Bayes with Intractable Likelihood (VBIL), when the likelihood can be estimated unbiasedly. VBIL transforms the problem of approximating the posterior $\pi(\theta)$ into a stochastic optimization problem using a noisy gradient. It is essential for the success of stochastic optimization algorithms to have a gradient estimator with a sufficiently small variance. Section 3 describes several techniques, including control variate and quasi-Monte Carlo, for variance reduction in estimating the gradient. This section also discusses the importance of the natural gradient (Amari, 1998), which takes into account the geometry of the variational distribution $q(\theta)$ being learned.

Unlike many VB algorithms that are derived on a model-by-model basis and require analytical computation of some model-based expectations, one of the main advantages of VBIL is that it can be applied to almost all statistical models without requiring an analytical solution to model-based expectations. The only requirement is that we are able to estimate the intractable likelihood unbiasedly. The VBIL methodology is therefore generic and widely applicable. As a by-product, VBIL provides an estimate of the marginal likelihood, which is useful for model choice.

There are several lines of work related to ours in terms of working with an intractable likelihood. Beaumont (2003) and Andrieu and Roberts (2009) show that Markov chain Monte Carlo simulation based on an unbiased estimator of the likelihood is still able to generate samples from the posterior. This method is known in the literature as Pseudo-Marginal Metropolis-Hasting (PMMH). More efficient variants of PMMH, called correlated PMMH and blockwise PMMH, have been proposed recently (Deligiannidis et al., 2015; Tran et al., 2016). Tran et al. (2013) show that importance sampling with the likelihood replaced by its unbiased estimator is still valid for estimating expectations with respect to the posterior, and name their method as Importance Sampling Squared (IS^2). Also, Duan and Fulop (2013) and Tran et al. (2014) use sequential Monte Carlo for inference based on an unbiased likelihood estimator. The main advantage of VBIL is that it is several orders of magnitude faster than these competitors.

Section 4 studies the link between the precision of the likelihood estimator to the variance of the VBIL estimator. This helps to understand how much accuracy is lost when working with an estimated likelihood compared to the case the likelihood is available. In this spirit, Pitt et al. (2012) and Tran et al. (2013) show that the asymptotic variance of PMMH and IS^2 estimators increases exponentially with the variance of the likelihood estimator. Therefore, it is critical for these methods to have a likelihood estimator that is accurate enough. They show that the variance of the likelihood estimator should be around 1 in order to minimize the computing time that is needed for the variance of PMMH/IS^2 estimators to have a fixed precision. For VBIL, we show that the asymptotic variance of VBIL estimators increases linearly with the variance of the likelihood estimator. The proposed methodology is therefore useful in cases when only highly variable estimates of the likelihood are available. We discuss such a situation in Section 5.1 where VBIL works well while its competitors fail.

Several interesting applications of VBIL are presented in Section 5. Section 5.1 shows the use of VBIL for generalized linear mixed models and demonstrates the high accuracy of VBIL compared to the existing VB algorithms. Section 5.2 applies VBIL to Bayesian inference in state space models and Section 5.3 shows how VBIL can be used for ABC. To the best of our knowledge, our paper is the first to use a VB method in the most general way for Bayesian inference in state space models and ABC. Another interesting application of VBIL is presented in Section 5.4, in which we illustrate that VBIL provides an attractive way to improve the accuracy of VB approximations of marginal posteriors.

2 Variational Bayes with an intractable likelihood

This section describes the basic form of the proposed VBIL algorithm, where an unbiased estimator of the likelihood is available. Denote by $\widehat{p}_N(y|\theta)$ an unbiased estimator of the likelihood $p(y|\theta)$. Here N is an algorithmic parameter relating to the precision in estimating the likelihood, such as the number of samples if the likelihood is estimated by importance sampling or the number of particles if the likelihood in state space models is estimated by a particle filter. Using the terminology in Pitt et al. (2012), we refer to N as the number of particles. Let $z = \log \widehat{p}_N(y|\theta) - \log p(y|\theta)$, so that $\widehat{p}_N(y|\theta) = p(y|\theta)e^z$, and denote by $g_N(z|\theta)$ the density of z. Note that z

is unknown as we do not know $\log p(y|\theta)$ and there is no need to compute z in practice, but, as will become clear shortly, it is very convenient to work with z. We sometimes write $\widehat{p}_N(y|\theta)$ as $\widehat{p}_N(y|\theta,z)$. We note that $\int e^z g_N(z|\theta)dz = 1$ because of the unbiasedness of the estimator $\widehat{p}_N(y|\theta)$. Define the following density on the extended space $\Theta \times \mathbb{R}$

$$\pi_N(\theta,z) = \frac{p(\theta)p(y|\theta)e^z g_N(z|\theta)}{p(y)} = \pi(\theta)e^z g_N(z|\theta).$$

This augmented density admits the posterior of interest $\pi(\theta)$ as its marginal. It is useful to work with $\pi_N(\theta,z)$ as the high-dimensional vector of random variables involved in estimating the likelihood is transformed into the scalar z. A direct approximation of $\pi_N(\theta,z)$ is $\widetilde{q}_{\lambda,N}(\theta,z) = q_\lambda(\theta)e^z g_N(z|\theta)$, where $q_\lambda(\theta)$ is the variational distribution with the variational parameter λ to be estimated, and then $q_\lambda(\theta)$ can be used as an approximation of $\pi(\theta)$. However, it turns out that it is impossible to estimate the gradient of the Kullback-Leibler divergence between $\widetilde{q}_{\lambda,N}(\theta,z)$ and $\pi_N(\theta,z)$ as this requires knowing z.

We propose instead to approximate $\pi_N(\theta,z)$ by $q_{\lambda,N}(\theta,z) = q_\lambda(\theta)g_N(z|\theta)$. This augmented density has the attractive features that $q_\lambda(\theta)$ is its marginal for θ and it is possible to estimate the gradient of the Kullback-Leibler divergence $\mathrm{KL}(\lambda)$ between $q_{\lambda,N}(\theta,z)$ and $\pi_N(\theta,z)$ (c.f. (2) below). Although $q_{\lambda,N}(\theta,z)$ does not provide a good approximation of the posterior marginal of z, the latter is not of interest to us. Furthermore, under Assumptions 1 and 2 given in Section 4, the minimization of $\mathrm{KL}(\lambda)$ is equivalent to the minimization of the KL divergence between $q_\lambda(\theta)$ and $\pi(\theta)$.

The Kullback-Leibler divergence between $q_{\lambda,N}(\theta,z)$ and $\pi_N(\theta,z)$ is

$$\mathrm{KL}(\lambda) = \int q_\lambda(\theta)g_N(z|\theta)\log\frac{q_\lambda(\theta)g_N(z|\theta)}{\pi_N(\theta,z)}dzd\theta, \tag{1}$$

where we omit to indicate dependence on N for notational convenience. The gradient of $\text{KL}(\lambda)$ is

$$\begin{aligned}\nabla_\lambda \text{KL}(\lambda) &= \nabla_\lambda \int q_\lambda(\theta) g_N(z|\theta) \log \frac{q_\lambda(\theta)}{p(\theta)\widehat{p}_N(y|\theta,z)} dz d\theta \\ &= \int \Big(\nabla_\lambda [q_\lambda(\theta)] g_N(z|\theta) \log \frac{q_\lambda(\theta)}{p(\theta)\widehat{p}_N(y|\theta,z)} + q_\lambda(\theta) g_N(z|\theta) \nabla_\lambda [\log q_\lambda(\theta)] \Big) dz d\theta \\ &= \int \Big(q_\lambda(\theta) g_N(z|\theta) \nabla_\lambda [\log q_\lambda(\theta)] \Big(\log q_\lambda(\theta) - \log(p(\theta)\widehat{p}_N(y|\theta,z)) \Big) \Big) dz d\theta \\ &= \mathbb{E}_{\theta \sim q_\lambda(\theta), z \sim g_N(z|\theta)} \Big(\nabla_\lambda [\log q_\lambda(\theta)] \Big(\log q_\lambda(\theta) - \log(p(\theta)\widehat{p}_N(y|\theta,z)) \Big) \Big). \qquad (2)\end{aligned}$$

Here, we have used the facts that $\nabla_\lambda[q_\lambda(\theta)] = q_\lambda(\theta)\nabla_\lambda[\log q_\lambda(\theta)]$ and that $\mathbb{E}(\nabla_\lambda[\log q_\lambda(\theta)]) = 0$. It follows from (2) that, by generating $\theta \sim q_\lambda(\theta)$ and $z \sim g_N(z|\theta)$, it is straightforward to obtain an unbiased estimator $\widehat{\nabla_\lambda \text{KL}}(\lambda)$ of the gradient $\nabla_\lambda \text{KL}(\lambda)$. Therefore, we can use stochastic optimization to optimize $\text{KL}(\lambda)$. We note that the unknown z is implicitly generated when the unbiased likelihood estimator $\widehat{p}_N(y|\theta) = \widehat{p}_N(y|\theta,z)$ is computed. In practice, z is never dealt with explicitly and it only plays a theoretical role in the mathematical derivations. The basic algorithm is as follows

Algorithm 1.
- *Initialize $\lambda^{(0)}$ and stop the following iteration if the stopping criterion is met.*
- *For $t = 0,1,...,$ compute $\lambda^{(t+1)} = \lambda^{(t)} - a_t \widehat{\nabla_\lambda KL}(\lambda^{(t)})$.*

We will refer to this algorithm as Variational Bayes with Intractable Likelihood (VBIL). The sequence $\{a_t\}$ should satisfy $a_t > 0$, $\sum_t a_t = \infty$ and $\sum_t a_t^2 < \infty$. We choose $a_t = 1/(1+t)$ in this paper. It is also possible to train a_t adaptively.

It is important to note that each iteration is parallelizable, as the gradient $\nabla_\lambda \text{KL}(\lambda)$ is estimated by independent samples from $q_{\lambda,N}(\theta,z)$.

2.1 Stopping criterion and marginal likelihood estimation

An easy-to-implement stopping rule is to stop the updating procedure if the change between $\lambda^{(t+1)}$ and $\lambda^{(t)}$, e.g. in terms of the Euclidean distance, is less than some threshold ϵ (Ranganath et al., 2014). However, it is difficult to select ϵ as such a distance depends on the scales and the length of the vector λ. It is easy to show that

$$\log p(y) = \int \log\left(\frac{p(\theta)\widehat{p}_N(y|\theta,z)}{q_\lambda(\theta)}\right) q_{\lambda,N}(\theta,z)dzd\theta + \text{KL}(\lambda) \geq \text{LB}(\lambda), \tag{3}$$

where

$$\begin{aligned}\text{LB}(\lambda) &= \int \log\left(\frac{p(\theta)\widehat{p}_N(y|\theta,z)}{q_\lambda(\theta)}\right) q_{\lambda,N}(\theta,z)dzd\theta \\ &= \mathbb{E}_{\theta,z}[\log p(\theta) - \log q_\lambda(\theta) + \log \widehat{p}_N(y|\theta,z)]\end{aligned} \tag{4}$$

is the lower bound on the log of the marginal likelihood $\log p(y)$. This lower bound after convergence can be used as an approximation to $\log p(y)$, which is useful for model selection purposes. The expectation of the first two terms in (4) can be computed analytically, while the last term can be estimated unbiasedly by samples from $q_{\lambda,N}(\theta,z)$. However, in our experience, estimating the entire expectation (4) based on samples from $q_{\lambda,N}(\theta,z)$ leads to a smaller variance. Denote by $\widehat{\text{LB}}(\lambda)$ the resulting unbiased estimate of $\text{LB}(\lambda)$. Although $\text{LB}(\lambda)$ is strictly non-decreasing over iterations, its sample estimate $\widehat{\text{LB}}(\lambda)$ might not be. To account for this, we suggest to stop the updating procedure if the change in an averaged value of the lower bounds over a window of M iterations, $\overline{\text{LB}}(\lambda_t) = (1/M)\sum_{k=1}^{M}\widehat{\text{LB}}(\lambda_{t-k+1})$, is less than some threshold ϵ. At convergence, the values $\text{LB}(\lambda_t)$ stay the same, therefore $\overline{\text{LB}}(\lambda_t)$ will average out the noise in $\widehat{\text{LB}}(\lambda_t)$ and is stable. Furthermore, we suggest to replace $\widehat{\text{LB}}(\lambda)$ by a scaled version of it, $\widehat{\text{LB}}(\lambda)/n$ with n the size of the dataset such as the number of observations. The scaled lower bound is more or less independent of the size of the dataset (c.f., Figure 3). We set $M=5$ and $\epsilon=10^{-5}$ in this paper.

3 Variance reduction and natural gradient

As is typical of stochastic optimization algorithms, the performance of Algorithm 1 depends greatly on the variance of the noisy gradient. This section describes several techniques for variance reduction.

3.1 Control variate

Denote $\widehat{h}(\theta,z)=\log\left(p(\theta)\widehat{p}_N(y|\theta,z)\right)$ for notational simplicity. Let $\theta_s\sim q_\lambda(\theta)$ and $z_s\sim g_N(z|\theta_s)$, $s=1,...,S$, be S samples from the variational distribution $q_{\lambda,N}(\theta,z)$. A naive estimator of the ith element of $\nabla_\lambda\text{KL}(\lambda)$ is

$$\widehat{\nabla_{\lambda_i}\text{KL}}(\lambda)^{\text{naive}} = \frac{1}{S}\sum_{s=1}^{S}\nabla_{\lambda_i}[\log q_\lambda(\theta_s)]\big(\log q_\lambda(\theta_s)-\widehat{h}(\theta_s,z_s)\big), \tag{5}$$

whose variance is often too large to be useful. For any number c_i, consider

$$\widehat{\nabla_{\lambda_i}\text{KL}}(\lambda) = \frac{1}{S}\sum_{s=1}^{S}\nabla_{\lambda_i}[\log q_\lambda(\theta_s)](\log q_\lambda(\theta_s)-\widehat{h}(\theta_s,z_s)-c_i), \tag{6}$$

which is still an unbiased estimator of $\nabla_{\lambda_i}\text{KL}(\lambda)$ since $\mathbb{E}(\nabla_\lambda[\log q_\lambda(\theta)])=0$, whose variance can be greatly reduced by an appropriate choice of c_i. Similar ideas are considered in the literature, see Paisley et al. (2012), Nott et al. (2012) and Ranganath et al.

(2014). The variance of $\widehat{\nabla_{\lambda_i}\mathrm{KL}}(\lambda)$ is

$$\begin{aligned}\mathbb{V}(\widehat{\nabla_{\lambda_i}\mathrm{KL}}(\lambda)) &= \frac{1}{S}\mathbb{V}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\big(\log q_\lambda(\theta)-\widehat{h}(\theta,z)-c_i\big)\Big)\\ &= \frac{1}{S}\mathbb{V}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\big(\log q_\lambda(\theta)-\widehat{h}(\theta,z)\big)\Big)\\ &\qquad -\frac{2c_i}{S}\mathrm{cov}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\big(\log q_\lambda(\theta)-\widehat{h}(\theta,z)\big),\nabla_{\lambda_i}[\log q_\lambda(\theta)]\Big)\\ &\qquad +\frac{c_i^2}{S}\mathbb{V}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\Big).\end{aligned}$$

The optimal c_i that minimizes this variance is

$$c_i=\mathrm{cov}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\big(\log q_\lambda(\theta)-\widehat{h}(\theta,z)\big),\nabla_{\lambda_i}[\log q_\lambda(\theta)]\Big)\Big/\mathbb{V}\Big(\nabla_{\lambda_i}[\log q_\lambda(\theta)]\Big). \quad (7)$$

Then

$$\mathbb{V}(\widehat{\nabla_{\lambda_i}\mathrm{KL}}(\lambda))=\mathbb{V}(\widehat{\nabla_{\lambda_i}\mathrm{KL}}(\lambda)^{\mathrm{naive}})(1-\rho_i^2)\le\mathbb{V}(\widehat{\nabla_{\lambda_i}\mathrm{KL}}(\lambda)^{\mathrm{naive}}),$$

where ρ_i is the correlation between $\nabla_{\lambda_i}[\log q_\lambda(\theta)]\big(\log q_\lambda(\theta)-\widehat{h}(\theta,z)\big)$ and $\nabla_{\lambda_i}[\log q_\lambda(\theta)]$. Often, ρ_i^2 is very close to 1.

We estimate the numbers c_i by samples $(\theta_s,z_s)\sim q_{\lambda,N}(\theta,z)$ as in (7). In order to ensure the unbiasedness of the gradient estimator, the samples used to estimate c_i must be independent of the samples used to estimate the gradient. In practice, the c_i can be updated sequentially as follows. At iteration t, we use the c_i computed in the previous iteration $t-1$, i.e. based on the samples from $q_{\lambda^{(t-1)},N}(\theta,z)$, to estimate the gradient $\widehat{\nabla_\lambda\mathrm{KL}}(\lambda^{(t)})$, which is estimated using new samples from $q_{\lambda^{(t)},N}(\theta,z)$. We then update the c_i using this new set of samples. By doing so, the unbiasedness is guaranteed while no extra samples are needed in updating the numbers c_i.

The gradient in the form of (2) can be written as a sum of two terms, where the first term $\mathbb{E}_{\theta\sim q_\lambda(\theta)}(\nabla_\lambda[\log q_\lambda(\theta)]\log q_\lambda(\theta))$ can be in most cases computed analytically.

However, as pointed out by a referee, this term should be estimated using the same samples of θ as we do in (6). Doing so helps to reduce the noises in estimating the gradient. This is because the first term plays the role of a control variate. This phenomenon is discussed in detail in Salimans and Knowles (2013).

3.2 Natural gradient

Intuitively, a different learning rate should be used for each scale in the gradient vector. That is, the traditional gradient vector $\nabla_\lambda \text{KL}(\lambda)$ should be multiplied by an appropriate scale matrix. It is well-known that the traditional gradient defined on the Euclidean space does not adequately capture the geometry of the variational distribution $q_\lambda(\theta)$ (Amari, 1998). A small Euclidean distance between λ and λ' does not necessarily mean a small Kullback-Leibler divergence between $q_\lambda(\theta)$ and $q_{\lambda'}(\theta)$. Amari (1998) defines the natural gradient as

$$\nabla_\lambda \text{KL}(\lambda)^{\text{natural}} = I_F(\lambda)^{-1} \nabla_\lambda \text{KL}(\lambda), \tag{8}$$

with $I_F(\lambda)$ the Fisher information matrix, and suggests using the natural gradient as an efficient alternative to the traditional gradient. See also Hoffman et al. (2013).

If the variational distribution $q_\lambda(\theta)$ has the exponential family form

$$q_\lambda(\theta) = \exp(T(\theta)'\lambda - Z(\lambda)), \tag{9}$$

with $T(\theta)$ the vector of sufficient statistics and λ the vector of natural parameters, then $I_F(\lambda) = \text{cov}_{q_\lambda}\big(T(\theta), T(\theta)\big)$ is computed analytically.

The use of the natural gradient in VB algorithms is considered, among others, by Honkela et al. (2010), Hoffman et al. (2013) and Salimans and Knowles (2013). We demonstrate the importance of the natural gradient using a simple example where the likelihood is available. We consider a model where the data $y_i{\sim}B(1{,}\theta)$ - the Bernoulli distribution with probability θ. We generate $n{=}200$ observations y_i from $B(1{,}\theta{=}0.3)$ and obtain $k{=}\sum_i y_i{=}57$. We use a uniform prior on θ. Then, the posterior $p(\theta|y)$ is Beta$(k{+}1{,}n{-}k{+}1)$. The variational distribution $q_\lambda(\theta)$ is chosen to be Beta$(\alpha{,}\beta)$ which belongs to the exponential family with the natural parameter $\lambda{=}(\alpha{,}\beta)'$. The Fisher information matrix $I_F(\lambda)$ is

$$I_F(\lambda)=\begin{bmatrix}\psi_1(\alpha)-\psi_1(\alpha+\beta) & \psi_1(\alpha+\beta)\\ \psi_1(\alpha+\beta) & \psi_1(\beta)-\psi_1(\alpha+\beta)\end{bmatrix},$$

where $\psi_1(x){=}\nabla_{xx}[\log\Gamma(x)]$ is the *trigamma* function. In this simple example, the gradient $\nabla_\lambda\mathrm{KL}(\lambda)$ in (2) can be computed analytically

$$\nabla_\lambda \mathrm{KL}(\lambda){=}I_F(\lambda)\lambda{-}H(\lambda)$$

with

$$H(\lambda)=\begin{bmatrix}k\psi_1(\alpha)-n\psi_1(\alpha+\beta)\\ (n-k)\psi_1(\beta)-n\psi_1(\alpha+\beta)\end{bmatrix}.$$

Using the traditional gradient, the update in Algorithm 1 is

$$\lambda^{(t+1)}=\lambda^{(t)}-a_t\Big(I_F(\lambda^{(t)})\lambda^{(t)}-H(\lambda^{(t)})\Big).$$

Using the natural gradient, the update is

$$\lambda^{(t+1)}=(1-a_t)\lambda^{(t)}+a_tI_F(\lambda^{(t)})^{-1}H(\lambda^{(t)}).$$

Figure 1 plots the densities of the exact posterior $\pi(\theta)$ and the variational distributions $q_\lambda(\theta)$ estimated by the VBIL using the traditional gradient and the natural gradient, with two different random initializations. The figure shows that the VBIL algorithm using the natural gradient is superior to using the traditional gradient. Furthermore, the VBIL algorithm based on the natural gradient is insensitive to the initial $\lambda^{(0)}$.

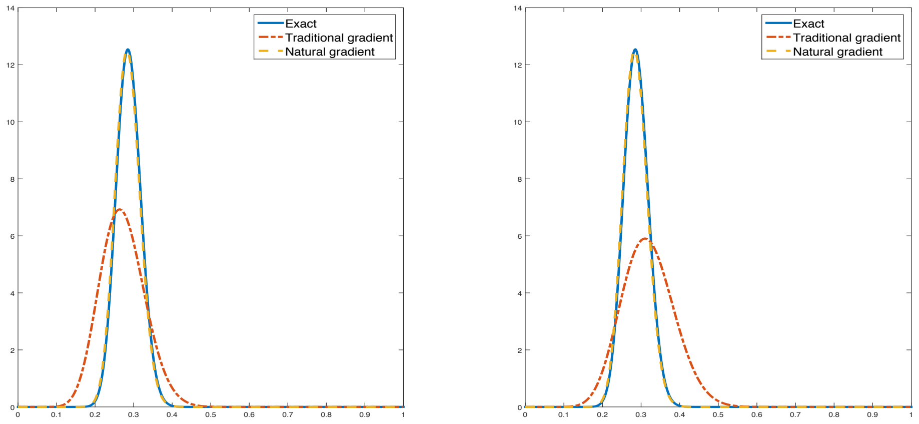

Figure 1: Plots of the densities of the exact posterior and the variational approximation estimates, at convergence, with two different starting values $\lambda^{(0)}$ at random.

3.3 Factorized variational distribution

Often, the variational distribution $q_\lambda(\theta)$ is factorized into K factors

$$q_\lambda(\theta) = q_{\lambda^{(1)}}(\theta^{(1)})...q_{\lambda^{(K)}}(\theta^{(K)}). \tag{10}$$

Then, each factor $q_{\lambda^{(k)}}(\theta^{(k)})$ is updated separately and the variance of the estimate of the corresponding gradient can be reduced. Salimans and Knowles (2013) and Ranganath et al. (2014) consider variance reduction using factorization. Denote by $\widehat{h}_k(\theta,z)$ the terms in $\widehat{h}(\theta,z)$ that involve only $\theta^{(k)}$ and z. From (2), and noting that

$\mathbb{E}_{\theta,z}(\nabla_{\lambda^{(k)}}[\log q_{\lambda^{(k)}}(\theta^{(k)})])=0$, the traditional gradient corresponding to factor k is

$$\nabla_{\lambda^{(k)}}\mathrm{KL}(\lambda)=\mathbb{E}_{\theta\sim q_{\lambda}(\theta),z\sim g_N(z|\theta)}\left(\nabla_{\lambda^{(k)}}[\log q_{\lambda^{(k)}}(\theta^{(k)})]\Big(\log q_{\lambda^{(k)}}(\theta^{(k)})-\widehat{h}_k(\theta,z)\Big)\right). \tag{11}$$

In the case $q_{\lambda^{(k)}}(\theta^{(k)})=\exp(T_k(\theta^{(k)})'\lambda^{(k)}-Z_k(\lambda^{(k)}))$ belongs to an exponential family, the natural gradient corresponding to factor k is

$$\nabla_{\lambda^{(k)}}\mathrm{KL}(\lambda)^{\text{natural}}=I_{F,k}(\lambda^{(k)})^{-1}\nabla_{\lambda^{(k)}}\mathrm{KL}(\lambda), \tag{12}$$

where $I_{F,k}(\lambda^{(k)})$ is the Fisher information matrix of distribution $q_{\lambda^{(k)}}(\theta^{(k)})$.

Estimating the gradient using (11) has less variation than using (2). Intuitively, this is because the variation due to terms not involving $\theta^{(k)}$ has been removed. This is also explained in Ranganath et al. (2014) as a Rao-Blackwellization effect.

3.4 Randomised quasi-Monte Carlo

Numerical integration using quasi-Monte Carlo (QMC) has been proven efficient in many applications. Instead of generating uniform random numbers $U(0,1)$ as in plain Monte Carlo methods, QMC generates deterministic sequences that are more evenly distributed in (0,1) in the sense that they minimise the so-called star-discrepancy. Dick and Pillichshammer (2010) provide an extensive background on QMC. It is shown that, in many cases, QMC integration achieves a better convergence rate than Monte Carlo integration. In this paper, we use randomised quasi-Monte Carlo (RQMC) as VBIL requires an unbiased estimator of the gradient. By introducing a random element into a QMC sequence, RQMC preserves the low-discrepancy property and, at the same time, leads to unbiased estimators (Owen, 1997; Dick and Pillichshammer, 2010).

Here, we use RQMC to sample $\theta \sim q_\lambda(\theta)$. This will help to reduce the variance of the noisy gradient if the dimension of θ is not too high. Of course, one can also use RQMC in the likelihood estimation, but given some time constraint we do not pursue this idea in this paper.

4 The effect of estimating the likelihood

This section studies the effect of the variance of the noisy likelihood on the VBIL estimators, and provides guidelines for selecting the number of particles N. A large N gives a precise likelihood estimate and therefore an accurate estimate of λ, but at a greater computational cost. A small N leads to a large variance of the likelihood estimator, so a larger number of iterations is needed for the procedure to settle down. It is therefore useful in practice to have some guidelines for selecting N.

In order to understand the effect of estimating the likelihood, we follow Pitt et al. (2012) and make the following assumption.

Assumption 1. *There is a function $\gamma^2(\theta) > 0$ such that $\mathbb{E}(z|\theta) = -\frac{\gamma^2(\theta)}{2N}$ and $\mathbb{V}(z|\theta) = \frac{\gamma^2(\theta)}{N}$.*

More precisely, Pitt et al. (2012) assume further that $z \sim \mathcal{N}(-\frac{\gamma^2(\theta)}{2N}, \frac{\gamma^2(\theta)}{N})$ in order to derive a theory for selecting an optimal N. This assumption is justified in Tran et al. (2013) and Doucet et al. (2015) making use of the unbiasedness of the likelihood estimate. The reason that the mean of z is $-\frac{1}{2}$ times its variance is because $\mathbb{E}(e^z) = 1$ in order for the likelihood estimator to be unbiased.

Assumption 2. *For a given $\sigma^2 > 0$, let N be a function of θ and σ^2 such that $\mathbb{V}(z|\theta) \equiv \sigma^2$, i.e. $N = N_{\sigma^2}(\theta) = \gamma^2(\theta)/\sigma^2$. Then $\mathbb{E}(z|\theta) = -\frac{\sigma^2}{2}$ and $\mathbb{V}(z|\theta) = \sigma^2$.*

Suppose that the equation $\nabla_\lambda \mathrm{KL}(\lambda) = 0$, with $\mathrm{KL}(\lambda)$ in (1), has the unique solution λ^*. Let $\widehat{\lambda}_M$ be the estimator of λ^* obtained by Algorithm 1 or 2 after M iterations, and $\widetilde{\lambda}_M$ be the corresponding estimator obtained when the exact likelihood is available. Denote $\zeta_*(\theta) = \nabla_\lambda [\log q_\lambda(\theta)]\big|_{\lambda=\lambda^*}$ and denote by $\mathbb{E}_*(.)$ and $\mathbb{V}_*(.)$ the expectation and variance operators with respect to $q_{\lambda^*}(\theta)$. For simplicity, we consider the case that λ is scalar; the case with a multivariate λ can be obtained using Theorem 5 of Sacks (1958). We obtain the following results whose proof is in the Appendix.

Theorem 1. *Suppose that Assumptions 1 and 2 are satisfied, and that the regularity conditions in Theorem 1 of Sacks (1958) hold.*

(i) Then,

$$\sqrt{M}(\widehat{\lambda}_M - \lambda^*) \xrightarrow{d} \mathcal{N}\Big(0, c_{\lambda^*}\mathbb{V}\big(\widehat{\nabla_\lambda \mathrm{KL}}(\lambda^*)\big)\Big), \ \textit{as } M \to \infty, \tag{13}$$

where c_{λ^*} is a positive constant that depends only on geometric properties of the function $\nabla_\lambda \mathrm{KL}(\lambda^*)$ and is independent of the random variables involving in estimating $\nabla_\lambda \mathrm{KL}(\lambda^*)$, i.e. c_{λ^*} is independent of σ^2.

(ii) Let $\sigma^2_{asym}(\widehat{\lambda}_M) = c_{\lambda^*}\mathbb{V}\big(\widehat{\nabla_\lambda \mathrm{KL}}(\lambda^*)\big)$ be the asymptotic variance of $\widehat{\lambda}_M$ as $M \to \infty$. Similarly, let $\sigma^2_{asym}(\widetilde{\lambda}_M)$ be the asymptotic variance of $\widetilde{\lambda}_M$. Then,

$$\sigma^2_{asym}(\widehat{\lambda}_M) = \sigma^2_{asym}(\widetilde{\lambda}_M) + \sigma^2 \tau(\lambda^*, S), \tag{14}$$

where $\tau(\lambda^*, S) = c_{\lambda^*}\mathbb{V}_*\big\{\zeta_*(\theta)\big\}/S$ if the noisy traditional gradient is used, and $\tau(\lambda^*, S) = c_{\lambda^*} I_F(\lambda^*)^{-1}\mathbb{V}_*\big\{\zeta_*(\theta)\big\} I_F(\lambda^*)^{-1}/S$ if the noisy natural gradient in (8) is used.

These results show that the variance of VBIL estimators increases linearly with σ^2. For PMMH and IS^2 estimators, Pitt et al. (2012) and Tran et al. (2013) show that their variances increase exponentially with σ^2. This means that VBIL is useful in cases where only a rough estimate of the likelihood is available, or it is expensive to obtain an accurate estimate of the likelihood.

We now discuss the issue of selecting σ^2. We note that under Assumption 2, N is tuned depending on θ as $N = N_{\sigma^2}(\theta) = \gamma^2(\theta)/\sigma^2$, so the time to compute the likelihood estimate $\widehat{p}_N(y|\theta)$ is proportional to $1/\sigma^2$. Then, Pitt et al. (2012) and Tran et al. (2013) show that, for the PMMH and IS^2 methods, the optimal σ^2 that gives an optimal trade-off between the CPU time and the variance of the estimators is 1. For VBIL, the computing time can be defined as

$$\mathrm{CT}(\sigma^2) = \frac{\sigma^2_{\mathrm{asym}}(\widehat{\lambda}_M)}{\sigma^2} = \frac{\sigma^2_{\mathrm{asym}}(\widetilde{\lambda}_M)}{\sigma^2} + \tau(\lambda^*, S), \tag{15}$$

where neither $\sigma^2_{\mathrm{asym}}(\widetilde{\lambda}_M)$ nor $\tau(\lambda^*,S)$ depends on σ^2. These results suggest that σ^2 should be set to a large value, as long as it is not too large for the stochastic search procedure in Algorithms 1 and 2 to converge.

5 Applications

5.1 Application to generalized linear mixed models and panel data models

Generalized linear mixed models (GLMM) (see, e.g. Fitzmaurice et al., 2011), also known as panel data models, use a vector of random effects α_i to account for the dependence between the observations $y_i = \{y_{ij}, j = 1,\dots,n_i\}$ measured on the same individual i. Given the random effects α_i, the conditional density $p(y_i|\theta,\alpha_i)$ belongs to an exponential family. The joint likelihood function of the model parameters θ

and the random effects $\alpha=(\alpha_1,...,\alpha_n)$, is tractable

$$p(y,\alpha|\theta)=\prod_{i=1}^{n}p(\alpha_i|\theta)p(y_i|\theta,\alpha_i).$$

Typically in the VB literature the joint posterior $p(\theta,\alpha|y)\propto p(\theta)p(y,\alpha|\theta)$ is approximated by a variational distribution of the form $q(\theta)q(\alpha)$, and then $q(\theta)$ is used as an approximation to the marginal posterior $p(\theta|y)$. For example, Tan and Nott (2013) take this approach but use partially non-centered parameterizations to reduce dependence between parameter blocks. Ormerod and Wand (2012) consider frequentist estimation of θ, but using VB methods to integrate out α. As discussed in the introduction, factorization of the VB distribution generally ignores the posterior dependence between θ and α, which often leads to underestimating the variance in the posterior distribution of θ. Below, we refer to such a VB method as classical VB.

The likelihood, $p(y|\theta)=\prod_{i=1}^{n}p(y_i|\theta)$ with $p(y_i|\theta)=\int p(y_i|\theta,\alpha_i)p(\alpha_i|\theta)d\alpha_i$ an integral over the random effects α_i, is in most cases analytically intractable but can be easily estimated unbiasedly using importance sampling. Let $h_i(\alpha_i|y,\theta)$ be an importance density for α_i. The integral $p(y_i|\theta)$ is estimated unbiasedly by

$$\widehat{p}_{N_i}(y_i|\theta)=\frac{1}{N_i}\sum_{j=1}^{N_i}w_i(\alpha_i^{(j)},\theta),\ w_i(\alpha_i^{(j)},\theta)=\frac{p(y_i|\alpha_i^{(j)},\theta)p(\alpha_i^{(j)}|\theta)}{h_i(\alpha_i^{(j)}|y,\theta)},\ \alpha_i^{(j)}\overset{iid}{\sim}h_i(\cdot|y,\theta). \quad (16)$$

It is possible to use different N_i for each $p(y_i|\theta)$. Hence, $\widehat{p}_N(y|\theta)=\prod_{i=1}^{n}\widehat{p}_{N_i}(y_i|\theta)$ is an unbiased estimator of the likelihood $p(y|\theta)$. The variance of $z=\log\widehat{p}_N(y|\theta)-\log p(y|\theta)$ is

$$\mathbb{V}(z|\theta)=\mathbb{V}(\log\widehat{p}_N(y|\theta))=\sum_{i=1}^{n}\mathbb{V}(\log\widehat{p}_{N_i}(y_i|\theta)), \quad (17)$$

which can be estimated by $\widehat{\mathbb{V}}(z|\theta)=\sum_{i=1}^n\widehat{\mathbb{V}}(\log\widehat{p}_{N_i}(y_i|\theta))$ with

$$\widehat{\mathbb{V}}(\log\widehat{p}_{N_i}(y_i|\theta)) = \frac{\widehat{\gamma}_i(\theta)}{N_i},\ \ \widehat{\gamma}_i(\theta) = \frac{N_i\sum_{j=1}^{N_i} w_i(\alpha_i^{(j)},\theta)^2}{\left(\sum_{j=1}^{N_i} w_i(\alpha_i^{(j)},\theta)\right)^2} - 1. \tag{18}$$

Given a fixed σ^2, it is therefore straightforward to target $\mathbb{V}(z|\theta)=\sigma^2$ by selecting N_i such that $\widehat{\mathbb{V}}(\log\widehat{p}_{N_i}(y_i|\theta))\approx\sigma^2/n$.

Six City data

We now illustrate the VBIL algorithm using the Six City data in Fitzmaurice and Laird (1993). The data consist of binary responses y_{ij} which indicate the wheezing status (1 if wheezing, 0 if not wheezing) of the ith child at time-point j, $i=1,...,537$ and $j=1,...,4$. Covariates are the age of the child at time-point j, centered at 9 years, and the maternal smoking status (0 or 1). We consider the following logistic regression model with a random intercept

$$\begin{aligned} p(y_{ij}|\beta,\alpha_i) &= \text{Binomial}(1,p_{ij}),\\ \text{logit}(p_{ij}) &= \beta_1+\beta_2\text{Age}_{ij}+\beta_3\text{Smoke}_{ij}+\alpha_i,\ \alpha_i\sim\mathcal{N}(0,\tau^2). \end{aligned}$$

The model parameters are $\theta=(\beta,\tau^2)$. We use a normal prior $\mathcal{N}(0,50I_3)$ for β and a Gamma(1,0.1) prior for τ^2.

We use the variational distribution $q_\lambda(\theta)=q(\beta)q(\tau^2)$, where $q(\beta)$ is a $d=3-$variate normal $\mathcal{N}(\mu,\Sigma)$ and $q(\tau^2)$ is an inverse gamma distribution. We then run Algorithm 2, see the Appendix for the detail, with $S=1000$ samples to estimate the gradient. The likelihood is estimated as in (16) with the natural sampler $h_i(\alpha_i|y,\theta)=p(\alpha_i|\theta)$, which is the normal distribution $\mathcal{N}(0,\tau^2)$ in our case. The σ^2 in Section 4 is set to 4, which on average requires $\bar{N}=\sum N_i/n=124$ particles. Using a larger σ^2 will lead to

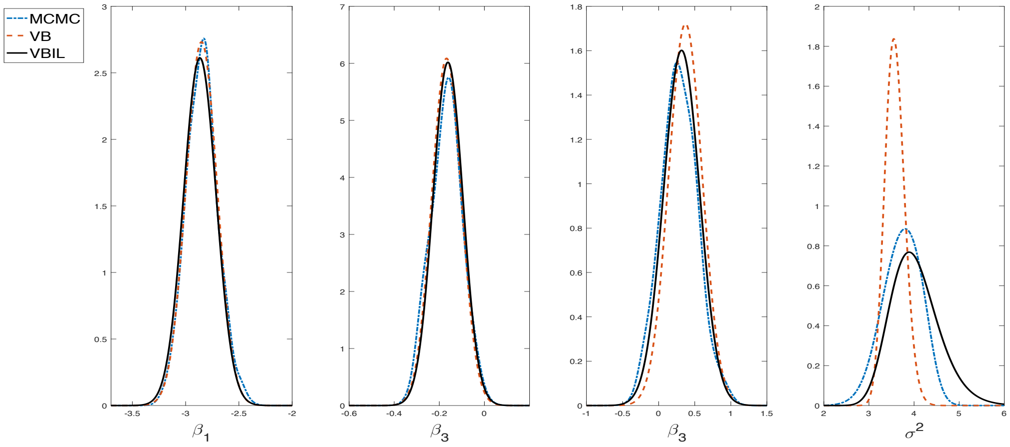

Figure 2: Application to GLMM: Six City data

too small N_i that makes the estimate in (18) unreliable. Figure 3(a) plots the scaled lower bounds over the iterations.

We compare the performance of the classical VB and VBIL algorithms to the pseudo-marginal MCMC simulation (Andrieu and Roberts, 2009). We set $\sigma^2 = 1$ as suggested in Pitt et al. (2012). The MCMC chain, based on the adaptive random walk Metropolis-Hastings algorithm in Haario et al. (2001), consists of 20000 iterates with another 10000 iterates used as burn-in.

Figure 2 plots the classical VB estimates (dashed line), MCMC estimates (dotted line) and the VBIL estimates (solid line) of the marginal posteriors $p(\beta_i|y)$ and $p(\tau^2|y)$. The MCMC density estimates are carried out using the kernel density estimation method based on the built-in Matlab function `ksdensity`. The figure shows that the VBIL estimates are very close to the MCMC estimates. The classical VB underestimates the posterior variance of τ^2 in this example. The clock times taken to run the VB, VBIL and MCMC procedures are 4, 2.9 and 505 minutes respectively. However, we note that the running time depends on many factors such as the programming language being used and the initialization of the procedures. All the

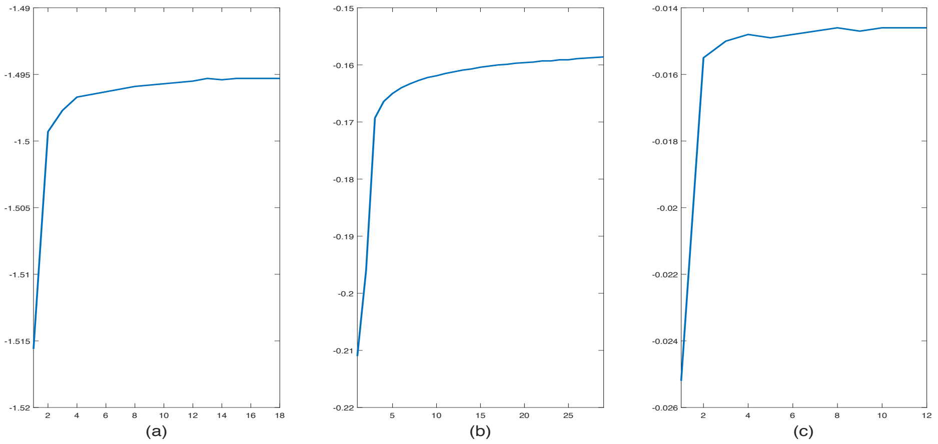

Figure 3: Plots of scaled lower bounds over the iterations: (a) Six City example, (b) state space example, (c) ABC example

examples in this paper are run on an Intel Core 16 i7 3.2GHz desktop supported by the Matlab Parallel Toolbox with 8 local processors. Obviously, the more processors we have, the faster the VBIL procerdure is.

Large data example

One of the main advantages of VBIL is its scalability, i.e. it is applicable in large data cases. This section describes a scenario where it is difficult to use the PMMH and IS^2 methods. Consider a large data case with a large number of panels n. From (17), for fixed N_i, the variance of the log-likelihood estimator $\mathbb{V}(z|\theta)$ increases linearly with n. Therefore, when n is large enough, the PMMH and IS^2 methods will not work in a practical sense, because $\mathbb{V}(z|\theta)$ can be very large (Flury and Shephard, 2011). In this GLMM setting, PMMH and IS^2 do not work when $\mathbb{V}(z|\theta)$ is as large as 6 or 7. One can decrease $\mathbb{V}(z|\theta)$ by increasing N_i, but this can be too computationally expensive to be practical.

We generate a data set of n=3000 from the following logistic model with a random intercept

$$\begin{aligned} p(y_{ij}|\beta,\alpha_i) &= \text{Binomial}(1,p_{ij}), \\ \text{logit}(p_{ij}) &= \beta_1+\beta_2 x_{ij}+\alpha_i,\ \ \alpha_i\sim N(0,\tau^2),\ \ i=1,...,n, j=1,...,n_i, \end{aligned} \tag{19}$$

with $\beta\!=\!(-1.5{,}2.5)'$, $\tau^2\!=\!1.5$, $n_i\!=\!5$, $x_{ij}\!\sim\! U(0{,}1)$. It takes, on average across different θ, 30 seconds to carry out each likelihood estimation with the numbers of particles N_i tuned to target $\mathbb{V}(z|\theta)\!=\!1$, which requires $\bar{N}\!=\!\sum N_i/n\!=\!3855$ particles. So if an optimal PMMH procedure was run on our computer to generate a chain of 30000 iterations as done in the Six City data example, it would take 10.4 days. We now run VBIL with σ^2 set to 30, which on average requires $\bar{N}\!=\!\sum N_i/n\!=\!187$ particles and 0.7 seconds for each likelihood estimation. The VBIL procedure stopped after 15 iterations with the clock time taken was 23 minutes. Figure 4 plots the variational approximations of the marginal posteriors, which are bell shaped as expected with a large dataset.

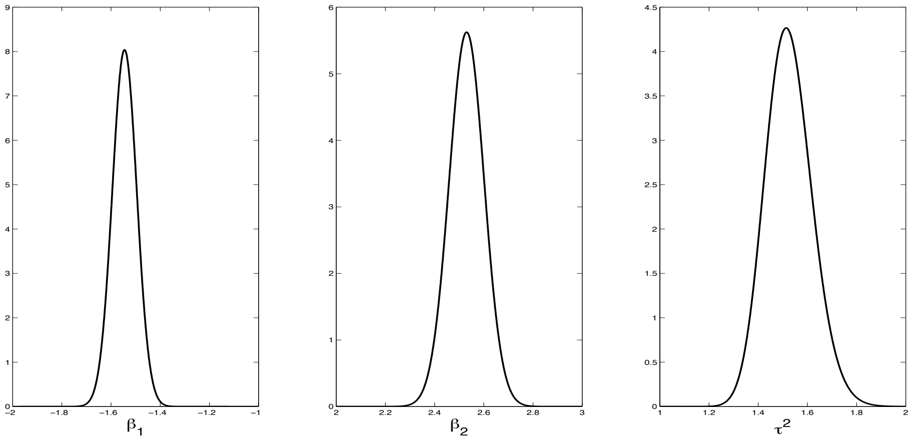

Figure 4: Application to GLMM: large data

5.2 Application to state space models

In state space models, the observations y_t are observed in time order. At time t, the distribution of y_t conditional on a state variable x_t is independently distributed as

$$y_t|x_t \sim g_t(y_t|x_t, \theta),$$

and the state variables $\{x_t\}_{t\geq 1}$ are a Markov chain with

$$x_1 \sim \mu_\theta(\cdot), \quad x_t|x_{t-1} \sim f_t(x_t|x_{t-1}, \theta).$$

The likelihood of the data $y = y_{1:T}$ is given by

$$p(y|\theta) = \int p(y|x, \theta)p(x|\theta)dx \tag{20}$$

with $x = x_{1:T}$ and

$$p(x|\theta) = \mu_\theta(x_1)\prod_{t=2}^{T} f_t(x_t|x_{t-1}, \theta), \;\; p(y|x, \theta) = \prod_{t=1}^{T} g_t(y_t|x_t, \theta).$$

Given a value of θ, the likelihood $p(y|\theta)$ can be unbiasedly estimated by an importance sampling estimator (Shephard and Pitt, 1997; Durbin and Koopman, 1997) or by a particle filter estimator (Del Moral, 2004; Pitt et al., 2012), $\widehat{p}_N(y|\theta)$, with N the number of particles.

An important example of state space models is the stochastic volatility (SV) model. The time series data y_t is modeled as

$$\begin{aligned} y_t &= \exp(x_t/2)w_t,\ w_t \sim \mathcal{N}(0,1), \\ x_t &= \mu + \phi(x_{t-1}-\mu) + \sigma v_t,\ x_1 \sim \mathcal{N}(\mu, \frac{\sigma^2}{1-\phi^2}),\ v_t \sim \mathcal{N}(0,1), \end{aligned}$$

with $\mu \in \mathbb{R}$, $\phi \in (-1,1)$ and $\sigma^2 > 0$. Let $\tau = (1+\phi)/2 \in (0,1)$; we will estimate τ but report results for ϕ. The model parameters are $\theta = (\mu,\tau,\sigma^2)$. We follow Kim et al. (1998) and use a normal prior $\mathcal{N}(0,10)$ for μ, a Beta prior $B(20,1.5)$ for τ and an inverse gamma IG(2.5,0.025) for σ^2.

To illustrate the VBIL algorithm for state space models, we analyze the weekday close exchange rates r_t for the Australian Dollar/US Dollar from 5/1/2010 to 31/12/2013. The data are available from the Reserve Bank of Australia. The data y_t is

$$y_t = 100\left(\log\frac{r_{t+1}}{r_t} - \frac{1}{T}\sum_{i=1}^{T}\log\frac{r_{i+1}}{r_i}\right),\ t = 1,...,T = 1001.$$

We use the variational distribution $q_\lambda(\theta) = q(\mu)q(\tau)q(\sigma^2)$, where $q(\mu)$ is $\mathcal{N}(\mu_\mu,\sigma_\mu^2)$, $q(\tau)$ is Beta(α_τ,β_τ) and $q(\sigma^2)$ is inverse gamma IG$(\alpha_{\sigma^2},\beta_{\sigma^2})$. We employ the constraint $\alpha_\tau > 1$ and $\beta_\tau > 1$ to make sure that $q(\tau)$ has a mode. The likelihood estimator $\widehat{p}_N(y|\theta)$ is computed by a basic particle filter. We then run the VBIL algorithm with $S = 1000$ samples, starting with $\mu_\mu = 0$, $\sigma_\mu^2 = 0.3$, $\alpha_\tau = 95$, $\beta_\tau = 5$, $\alpha_{\sigma^2} = 11$, $\beta_{\sigma^2} = 1$. This initial point is set so that the initial mean values of μ, ϕ and σ^2 are 0, 0.9 and 0.1 respectively, which is pretty far away from the posterior means; see Figure 5. The VBIL algorithm stops after 28 iterations. Figure 3(b) plots the scaled lower bounds over the iterations.

The VBIL is compared to pseudo-marginal MCMC simulation, based on an adaptive random walk Metropolis-Hastings algorithm, with 100,000 iterations starting from the same values $\mu = 0$, $\tau = 0.95$ and $\sigma^2 = 0.1$. The number of particles used in MCMC is fixed at $N = 300$, so that $\mathbb{V}(\widehat{p}_N(y|\bar{\theta})) \approx 1$ at the initial value $\bar{\theta} = (0{,}0.95{,}0.1)$. The number of particles used in VBIL is fixed at $N = 100$ as the use of randomised QMC for generating θ helps reduce greatly the variance in estimating the gradient. We fix N in this example as it is difficult to estimate the variance of log-likelihood estimates obtained by the particle filter.

Figure 5 plots the MCMC estimates (dotted line) and the VBIL estimates (solid line) of the marginal posteriors. The figure shows that the VBIL estimates are close to the MCMC estimates but consume significantly less computational resources. The CPU times taken to run the VBIL and MCMC procedures are 0.7 and 28 minutes respectively.

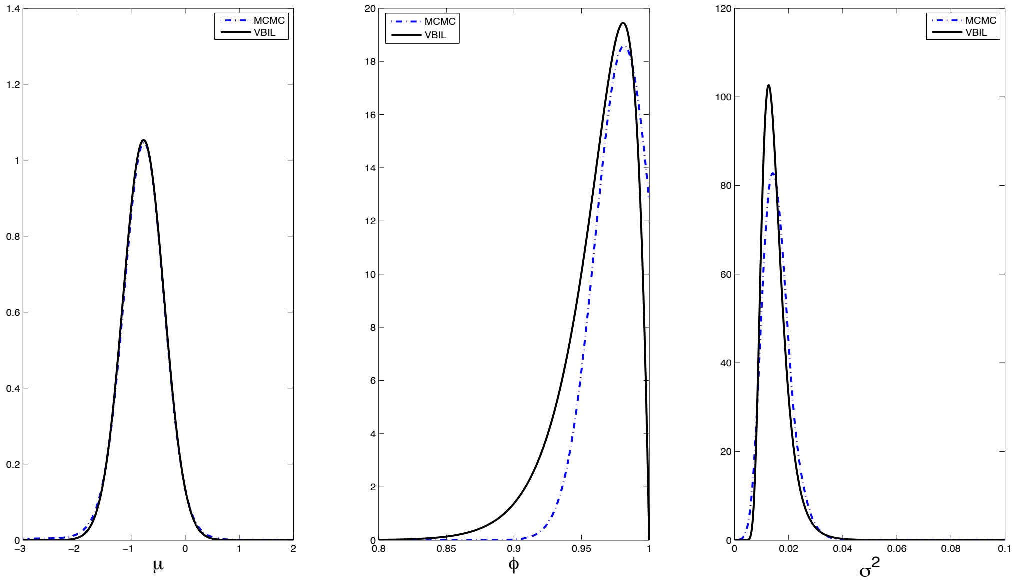

Figure 5: Application to state space models: Exchange rate data

5.3 Application to ABC

In many modern applications, such as in genetics (Tavare et al., 1997), we either cannot evaluate the likelihood $p(y|\theta)$ pointwise or do not wish to do so, but we can sample from it, i.e. we can simulate $y' \sim p(\cdot|\theta)$. Approximate Bayesian computation (Tavare et al., 1997) approximates the likelihood by

$$p_{\text{LF}}(y|\theta) = \int K_\epsilon(S(y'), S(y))p(y'|\theta)dy', \tag{21}$$

where $K_\epsilon(.,.)$ is a kernel with the bandwidth ϵ and $S(.)$ is a vector of summary statistics. Inference is then based on the approximate posterior $p_{\text{ABC}}(\theta|y) \propto p(\theta)p_{\text{LF}}(y|\theta)$. Because the likelihood-free function $p_{\text{LF}}(y|\theta)$ can be unbiasedly estimated by

$$\widehat{p}_N^{\text{LF}}(y|\theta) = \frac{1}{N}\sum_{i=1}^{N} K_\epsilon(S(y^{[i]}), S(y)), \;\; y^{[i]} \overset{iid}{\sim} p(\cdot|\theta),$$

it is straightforward to use the VBIL algorithm to approximate $p_{\text{ABC}}(\theta|y)$.

We illustrate the application of the VBIL algorithm to ABC by using it to fit an α-stable distribution. α-stable distributions (Nolan, 2007) are a class of heavy-tailed distributions used in many statistical applications. An α-stable distribution $\mathcal{S}(\alpha,\beta,\gamma,\delta)$ is parameterized by the stability parameter $\alpha \in (0,2)$, skewness $\beta \in (-1,1)$, scale $\gamma > 0$ and location $\delta \in \mathbb{R}$. The main difficulty when working with α-stable distributions is that they do not have closed form densities, which makes it difficult to do inference. However, as it is easy to sample from an α-stable distribution, one can use ABC techniques for Bayesian inference (Peters et al., 2012). We illustrate in this example that VBIL provides an efficient approach for fitting α-stable distributions.

We generate a data set y with $n = 500$ observations from a univariate α-stable distribution $\mathcal{S}(1.5,0.5,1,0)$. Let $\widehat{q}_p$ be the pth quantile of a pseudo-data set y' generated from $\mathcal{S}(\alpha,\beta,\gamma,\delta)$. We follow Peters et al. (2012) and use the summary statistics $S(y')=(\widehat{v}_\alpha,\widehat{v}_\beta,\widehat{v}_\gamma,\widehat{v}_\delta)$ with

$$\widehat{v}_\alpha=\frac{\widehat{q}_{0.95}-\widehat{q}_{0.05}}{\widehat{q}_{0.75}-\widehat{q}_{0.25}},\ \widehat{v}_\beta=\frac{\widehat{q}_{0.95}+\widehat{q}_{0.05}-2\widehat{q}_{0.5}}{\widehat{q}_{0.95}-\widehat{q}_{0.05}},\ \widehat{v}_\gamma=\frac{\widehat{q}_{0.75}-\widehat{q}_{0.25}}{\gamma},\ \widehat{v}_\delta=\frac{1}{n}\sum_{i=1}^n y_i'.$$

For the observed data y, the parameter γ in $\widehat{v}_\gamma$ is estimated using McCulloch's method (McCulloch, 1986). As the parameterization is discontinuous at $\alpha=1$, resulting in poor estimates of the summary statistics, we consider the case with $\alpha>1$ and restrict the support of α to the interval (1.1,2) as in Peters et al. (2012).

We reparameterize

$$\widetilde{\alpha}=\log(\frac{\alpha-1.1}{2-\alpha})\in\mathbb{R},\ \ \widetilde{\beta}=\log(\frac{\beta+1}{1-\beta})\in\mathbb{R},\ \ \widetilde{\gamma}=\log(\gamma)\in\mathbb{R},\ \ \widetilde{\delta}=\delta\in\mathbb{R},$$

and estimate $\widetilde{\theta}=(\widetilde{\alpha},\widetilde{\beta},\widetilde{\gamma},\widetilde{\delta})$ but report the results for $(\alpha,\beta,\gamma,\delta)$. We use a normal prior $\widetilde{\theta}\sim\mathcal{N}(0,100I_4)$ and approximate the posterior $p(\widetilde{\theta}|y)$ by a normal variational distribution $q_\lambda(\widetilde{\theta})=N(\mu_{\widetilde{\theta}},\Sigma_{\widetilde{\theta}})$. One can work with the original parameterization $(\alpha,\beta,\gamma,\delta)$ and use some form of factorization $q(\alpha)q(\beta)q(\gamma)q(\delta)$. We choose to work with $\widetilde{\theta}$ to account for the posterior dependence between the parameters. This also illustrates the flexibility of the VBIL method in the sense that it can be used without requiring factorization.

We use the Gaussian kernel with covariance matrix $0.01I_4$ for the likelihood-free $p_{\text{LF}}(y|\theta)$ in (21). The VBIL is compared to pseudo-marginal Metropolis-Hastings methods with 20,000 iterations after 5000 burnins. For the standard PMMH (Andrieu and Roberts, 2009), the number of pseudo-data sets $N=20$ is selected set after some trials in order to have a well-mixing chain. Efficient versions of PMMH has been proposed recently, which are more tolerant of noise in the likelihood estimates. Here we compare VBIL to the blockwise PMMH method of Tran et al. (2016). For the blockwise PMMH, we set $N = 5$. We also use this value of N in VBIL. Table 1 shows the VBIL and MCMC estimates, and the CPU times. As shown, VBIL is orders of magnitude faster than MCMC in this example. Figure 3(c) plots the scaled lower bounds over the iterations.

	True	Standard PMMH	Blockwise PMMH	VBIL
α	1.5	1.57 (0.15)	1.58 (0.14)	1.57 (0.11)
β	0.5	0.46 (0.21)	0.45 (0.21)	0.48 (0.16)
γ	1	1.04 (0.12)	1.04 (0.12)	1.02 (0.12)
δ	0	-0.08 (0.21)	-0.09 (0.18)	-0.08 (0.14)
CPU time (min)		12.56	7.62	0.12

Table 1: ABC example: Standard PMMH, blockwise PMMH and VBIL estimates of α, β, γ and δ. The numbers in brackets are estimates of the posterior standard deviations.

5.4 Using VBIL to improve marginal posterior estimates

A drawback of VB methods in general is that the factorization assumption as in (10) ignores the posterior dependence between the factors, which might lead to poor approximations of the posterior variances (Neville et al., 2014). We now show how the VBIL algorithm can be used to help overcome this problem.

Suppose that we would like to have a highly accurate VB approximation to the marginal posterior $p(\theta^{(j)}|y)$. We restrict ourselves to the case with a tractable likelihood for simplicity, but the following discussion also applies when the likelihood is intractable. The likelihood of $\theta^{(j)}$,

$$p(y|\theta^{(j)}) = \int p(\theta^{(\setminus j)}|\theta^{(j)})p(y|\theta^{(1)},...,\theta^{(K)})d\theta^{(\setminus j)}, \tag{22}$$

with $\theta^{(\setminus j)} = (\theta^{(1)},...,\theta^{(j-1)},\theta^{(j+1)},...,\theta^{(K)})$, is in general intractable but can be estimated unbiasedly. Let $q(\theta^{(\setminus j)})$ be an approximation to the marginal posterior $p(\theta^{(\setminus j)}|y)$

resulting from a classical VB method that uses the factorization (10). The integral in (22) can be estimated unbiasedly using importance sampling with the proposal density $q(\theta^{(\setminus j)})$ or a tail-flattened version of it. This is accurate enough in practice because VBIL does not require a very accurate estimate of $p(y|\theta^{(j)})$ as discussed in Section 4. The VBIL algorithm can then be used to approximate the marginal posterior $p(\theta^{(j)}|y)$ directly with $\theta^{(\setminus j)}$ integrated out. The resulting approximation is often highly accurate as the dependence between $\theta^{(j)}$ and $\theta^{(\setminus j)}$ is taken into account.

A formal justification is as follows. We use the notation as in (10) and write $\lambda=(\lambda^{(j)},\lambda^{(\setminus j)})$. Suppose that we estimate the marginal posterior of $\lambda^{(j)}$ by $q_{\lambda^{(j)}}(\theta^{(j)})$ which belongs to a family $\mathcal{F}=\{q_{\lambda^{(j)}}(\theta^{(j)}),\lambda^{(j)}\in\Lambda\}$. VBIL proceeds by minimizing

$$\mathrm{KL}_j(\lambda^{(j)}) = \int q_{\lambda^{(j)}}(\theta^{(j)}) \log \frac{q_{\lambda^{(j)}}(\theta^{(j)})}{p(\theta^{(j)}|y)} d\theta^{(j)}$$

over $\lambda^{(j)}\in\Lambda$. Let $\lambda_*^{(j)}$ be the VBIL estimator. Under Assumptions 1 and 2 or when the number of samples N used to estimate (22) is large enough, $\lambda_*^{(j)}$ is guaranteed to be a minimizer of $\mathrm{KL}_j(\lambda^{(j)})$. Assume further that $\mathrm{KL}_j(\lambda^{(j)})$ is convex, then

$$\mathrm{KL}_j(\lambda_*^{(j)}) \leq \mathrm{KL}_j(\lambda^{(j)}) \;\text{ for all }\; \lambda^{(j)} \in \Lambda. \tag{23}$$

If we use a VB procedure with a factorization of the form $q_\lambda(\theta)=q_{\lambda^{(j)}}(\theta^{(j)})q_{\lambda^{(\setminus j)}}(\theta^{(\setminus j)})$ where $q_{\lambda^{(j)}}(\theta^{(j)})$ belongs to the same family $\mathcal{F}$, then VB proceeds by minimizing the

KL divergence

$$\begin{aligned}
\mathrm{KL}(\lambda^{(j)},\lambda^{(\backslash j)}) &= \int q_{\lambda^{(j)}}(\theta^{(j)})q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})\log\frac{q_{\lambda^{(j)}}(\theta^{(j)})q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})}{p(\theta^{(j)},\theta^{(\backslash j)}|y)}d\theta^{(j)}d\theta^{(\backslash j)}\\
&= \int q_{\lambda^{(j)}}(\theta^{(j)})\log\frac{q_{\lambda^{(j)}}(\theta^{(j)})}{p(\theta^{(j)}|y)}d\theta^{(j)}\\
&\qquad + \int q_{\lambda^{(j)}}(\theta^{(j)})\int q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})\log\frac{q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})}{p(\theta^{(\backslash j)}|\theta^{(j)},y)}d\theta^{(\backslash j)}d\theta^{(j)}\\
&= \mathrm{KL}_j(\lambda^{(j)}) + \int q_{\lambda^{(j)}}(\theta^{(j)})\int q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})\log\frac{q_{\lambda^{(\backslash j)}}(\theta^{(\backslash j)})}{p(\theta^{(\backslash j)}|\theta^{(j)},y)}d\theta^{(\backslash j)}d\theta^{(j)} \quad (24)
\end{aligned}$$

Let $(\widetilde{\lambda}^{(j)},\widetilde{\lambda}^{(\backslash j)})$ be a minimizer of (24). Because of the decomposition in (24), the estimator $\widetilde{\lambda}^{(j)}$ is not necessarily the minimizer of $\mathrm{KL}_j(\lambda^{(j)})$. From (23),

$$\mathrm{KL}_j(\lambda_*^{(j)}) \le \mathrm{KL}_j(\widetilde{\lambda}^{(j)}). \quad (25)$$

So the VBIL estimator $\lambda_*^{(j)}$ is no worse than the factorization-based VB estimator $\widetilde{\lambda}^{(j)}$ in terms of KL divergence.

We illustrate this application by generating $n=100$ observations from a univariate mixture of two normals

$$p(x) = \omega\mathcal{N}(x|\mu_1,\sigma_1^2) + (1-\omega)\mathcal{N}(x|\mu_2,\sigma_2^2)$$

with $\omega=0.3$, $\mu_1=-3$, $\mu_2=3$, $\sigma_1^2=2$ and $\sigma_2^2=3$. Suppose that we are interested in getting an accurate variational approximation of the posterior $p(\omega|y)$. Getting an accurate estimate of w is often more challenging than the other parameters. We use diffuse priors $\omega\sim U(0{,}1)$, $\mu_1\sim\mathcal{N}(0{,}100)$, $\mu_2\sim\mathcal{N}(0{,}100)$, $\sigma_1^2\sim(\sigma_1^2)^{-1}$ and $\sigma_2^2\sim(\sigma_2^2)^{-1}$, and run VBIL to approximate $p(\omega|y)$ by a Beta distribution. We use the VB algorithm of McGrory and Titterington (2007), in which the variational distribution is factorized as $q(\omega)q(\sigma_1^2,\sigma_2^2)q(\mu_1,\mu_2|\sigma_1^2,\sigma_2^2)$, to design the proposal density to obtain an importance sampling estimator of $p(y|\omega)$.

Figure 6 plots the McGrory-Titterington estimate (dashed line) and VBIL estimate (solid line) of the posterior $p(\omega|y)$. As shown, the VBIL estimate has heavier tails than the VB estimate. By (25), it follows that the difference between the two estimates gives an indication of the extent to which the McGrory-Titterington estimate is suboptimal. This example shows that the VBIL method provides an attractive way to obtain accurate approximation of marginal posteriors.

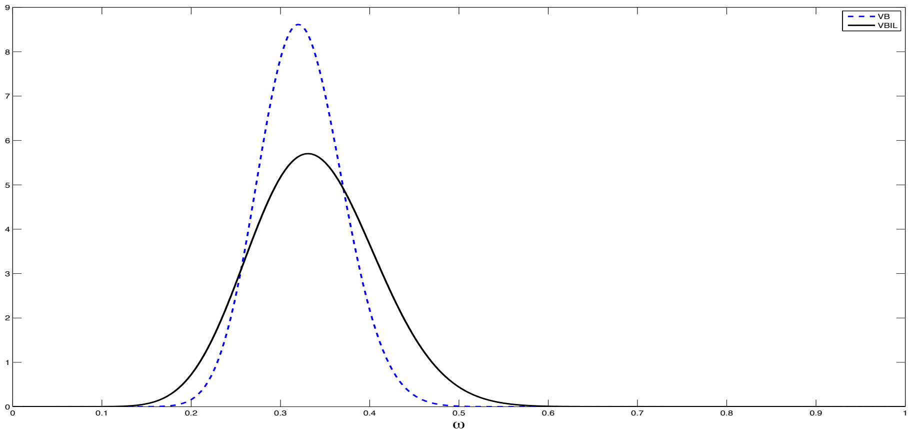

Figure 6: Plots the VB (dashed line) and VBIL estimates (solid line) of the posterior $p(\omega|y)$.

6 Conclusion

We have proposed the VBIL, a useful VB algorithm for Bayesian inference in statistical modeling where the likelihood is intractable. The method makes it possible to do inference in statistical models using VB in some situations where that was previously impossible. The main advantage of VBIL over its competitors, such as PMMH and IS^2, is its scalability. We show in the examples that VBIL is several orders of magnitude faster than these existing methods.

Acknowledgement

The research of Tran and Kohn was partially supported by the ARC COE grant CE140100049. Nott's research was supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112).

Appendix

Proof of Theorem 1. (i) Under Assumptions 1 and 2, we have that

$$\text{KL}(\lambda) = \text{KL}(q_\lambda \| \pi) - \int q_\lambda(\theta)\mathbb{E}(z|\theta)d\theta = \text{KL}(q_\lambda \| \pi) + \frac{\sigma^2}{2}, \tag{26}$$

where $\text{KL}(q_\lambda \| \pi)$ is the Kullback-Leibler divergence between the variational distribution $q_\lambda(\theta)$ and the posterior $\pi(\theta)$. So, $\nabla_\lambda \text{KL}(\lambda) = \nabla_\lambda \text{KL}(q_\lambda \| \pi)$ is independent of σ^2, and minimizing $\text{KL}(\lambda)$ with respect to λ is equivalent to minimizing $\text{KL}(q_\lambda \| \pi)$. Algorithm 1 and 2 are the Robbins-Monro procedure for finding the root λ^* of the equation $\nabla_\lambda \text{KL}(q_\lambda \| \pi) = 0$. Then, (13) follows from Theorem 1 of Sacks (1958) with the constant c_{λ^*} independent of σ^2.

(ii) Denote $\widehat{h}(\theta,z) = \log(p(\theta)\widehat{p}_N(y|\theta,z)) = \log(p(\theta)p(y|\theta)) + z = h(\theta) + z$. We consider the case with the noisy traditional gradient in (6); the proof for the other cases is similar. We denote by $\widetilde{\nabla_\lambda \text{KL}}(\lambda^*)$ the noisy gradient obtained when the likelihood is available. Then, noting that $\mathbb{E}_*(\zeta_*(\theta)) = 0$, the constant c in (7) is

$$c = \frac{\mathbb{E}_{\theta,z}\{\zeta_*(\theta)^2(\log q_{\lambda^*}(\theta) - h(\theta) - z)\}}{\mathbb{E}_*\big\{\zeta_*(\theta)^2\big\}} = \frac{\mathbb{E}_*\{\zeta_*(\theta)^2(\log q_{\lambda^*}(\theta) - h(\theta))\}}{\mathbb{E}_*\big\{\zeta_*(\theta)^2\big\}} + \frac{\sigma^2}{2} = \widetilde{c} + \frac{\sigma^2}{2}.$$

We note that $\widetilde{c}$ is the control variate constant we would use to compute $\widetilde{\nabla_\lambda\text{KL}}(\lambda^*)$ if the likelihood was known.

$$\begin{aligned}
\mathbb{V}\big(\widehat{\nabla_\lambda\text{KL}}(\lambda^*)\big) &= \frac{1}{S}\mathbb{V}_{\theta,z}\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta)-h(\theta)-z-c)\Big\}\\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\}+\frac{1}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta)-h(\theta)+\frac{\sigma^2}{2}-c)\Big\}\\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\}+\frac{1}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)(\log q_{\lambda^*}(\theta)-h(\theta)-\widetilde{c})\Big\}\\
&= \frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\}+\mathbb{V}\big(\widetilde{\nabla_\lambda\text{KL}}(\lambda^*)\big).
\end{aligned}$$

Therefore,

$$\sigma^2_{\text{asym}}(\widehat{\lambda}_M)=c_{\lambda^*}\mathbb{V}\big(\widehat{\nabla_\lambda\text{KL}}(\lambda^*)\big)=\sigma^2_{\text{asym}}(\widetilde{\lambda}_M)+c_{\lambda^*}\frac{\sigma^2}{S}\mathbb{V}_*\Big\{\zeta_*(\theta)\Big\}.$$

□

Derivation for Section 5.1

The density of the $d-$variate normal $\mathcal{N}(\mu,\Sigma)$ is

$$q(\beta)=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\Big(-\frac{1}{2}(\beta-\mu)'\Sigma^{-1}(\beta-\mu)\Big).$$

A simplified form of the inverse Fisher matrix for a multivariate normal under the natural parameterization is given in Wand (2014). For a $d\times d$ matrix A, denote by vec(A) the d^2-vector obtained by stacking the columns of A, by vech(A) the $\frac{1}{2}d(d+1)$-vector obtained by stacking the columns of the lower triangular part of A. The duplication matrix of order d, D_d, is the $d^2\times\frac{1}{2}d(d+1)$ matrix of zeros and ones such that for any symmetric matrix A

$$D_d\text{vech}(A) = \text{vec}(A).$$

Let $D_d^+ = (D_d' D_d)^{-1} D_d'$ be the Moore-Penrose inverse of D_d, and vec^{-1} be the inverse of the operator vec. Then, the exponential family form of the normal distribution $q(\beta)$ is $q(\beta) = \exp(T(\beta)'\lambda - Z(\lambda))$ with

$$T(\beta) = \begin{bmatrix} \beta \\ \text{vech}(\beta\beta') \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \end{bmatrix} = \begin{bmatrix} \Sigma^{-1}\mu \\ -\frac{1}{2} D_d' \text{vec}(\Sigma^{-1}) \end{bmatrix}. \tag{27}$$

The usual mean and variance parameterization is

$$\begin{cases} \mu = -\frac{1}{2}\left\{\text{vec}^{-1}(D_d^{+\prime}\lambda_2)\right\}^{-1}\lambda_1 \\ \Sigma = -\frac{1}{2}\left\{\text{vec}^{-1}(D_d^{+\prime}\lambda_2)\right\}^{-1}. \end{cases}$$

Wand (2014) derives the following very useful formula

$$I_F(\lambda)^{-1} = \begin{bmatrix} \Sigma^{-1} + M'S^{-1}M & -M'S^{-1} \\ -S^{-1}M & S^{-1} \end{bmatrix}, \tag{28}$$

with

$$M = 2D_d^+(\mu \otimes I_d) \text{ and } S = 2D_d^+(\Sigma \otimes \Sigma)D_d^{+\prime},$$

where $\otimes$ is the Kronecker product and I_d the identity matrix of order d. The gradient $\nabla_\lambda[\log q(\beta)]$ is

$$\nabla_\lambda[\log q(\beta)] = \begin{bmatrix} \beta - \mu \\ \text{vech}(\beta\beta' - \Sigma - \mu\mu') \end{bmatrix}. \tag{29}$$

For the inverse gamma distribution $q(\tau^2)$ with density

$$q(\tau^2) = \frac{a^b}{\Gamma(a)}(\tau^2)^{-1-a}\exp(-b/\tau^2),$$

the natural parameters are (a,b). The Fisher information matrix for the inverse gamma is

$$I_F(a,b) = \begin{pmatrix} \nabla_{aa}[\log\Gamma(a)] & -1/b \\ -1/b & a/b^2 \end{pmatrix}.$$

and the gradient

$$\begin{aligned} \nabla_a[\log q_\lambda(\theta)] &= -\log(\tau^2) + \log(b) - \nabla_a[\log\Gamma(a)] \\ \nabla_b[\log q_\lambda(\theta)] &= -\frac{1}{\tau^2} + \frac{a}{b}. \end{aligned}$$

References

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 72:1–33.

Andrieu, C. and Roberts, G. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics*, 37:697–725.

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, pages 21–30.

Barthelme, S. and Chopin, N. (2014). Expectation propagation for likelihood-free inference. *Journal of the American Statistical Association*, 109(505):315–333.

Beaumont, M. A. (2003). Estimation of population growth or decline in genetically monitored populations. *Genetics*, 164:1139–1160.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning.* New York: Springer.

Del Moral, P. (2004). *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications.* Springer, New York.

Deligiannidis, G., Doucet, A., and Pitt, M. (2015). The correlated pseudo-marginal method. Technical report. arXiv:1511.04992v3.

Dick, J. and Pillichshammer, F. (2010). *Digital nets and sequences. Discrepancy theory and quasi-Monte Carlo integration.* Cambridge University Press, Cambridge.

Doucet, A., Pitt, M. K., Deligiannidis, G., and Kohn, R. (2015). Efficient implementation of Markov chain Monte Carlo when using an unbiased likelihood estimator. arXiv:1210.1871.

Duan, J.-C. and Fulop, A. (2013). Density-tempered marginalized sequential Monte Carlo sampler. Technical report, National University of Singapore. Available at http://dx.doi.org/10.2139/ssrn.1837772.

Durbin, J. and Koopman, S. J. (1997). Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84:669–684.

Durbin, J. and Koopman, S. J. (2001). *Time Series Analysis by State Space Methods.* Oxford University Press.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analysing longitudinal binary responses. *Biometrika*, 80(1):141–151.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2011). *Applied Longitudinal Analysis*. John Wiley & Sons, Ltd, New Jersey, 2nd edition.

Flury, T. and Shephard, N. (2011). Bayesian inference based only on simulated likelihood: Particle filter analysis of dynamic economic models. *Econometric Theory*, 1:1–24.

Ghahramani, Z. and Hinton, G. E. (2000). Variational learning for switching state-space models. *Neural computation*, 12(4):831–864.

Haario, H., Saksman, E., and Tamminen, J. (2001). An adaptive Metropolis algorithm. *Bernoulli*, 7:223–242.

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013). Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347.

Honkela, A., Raiko, T., Kuusela, M., Tornio, M., and Karhunen, J. (2010). Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes. *Journal of Machine Learning Research*, 11:3235–3268.

Kim, S., Shephard, N., and Chib, S. (1998). Stochastic volatility: Likelihood inference and comparison with arch models. *The Review of Economic Studies*, 65(3):361–393.

Marin, J.-M., Pudlo, P., Robert, C., and Ryder, R. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180.

McCulloch, J. H. (1986). Simple consistent estimators of stable distribution parameters. *Communications in Statistics - Simulation and Computation*, 15(4):1109–1136.

McGrory, C. and Titterington, D. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics & Data Analysis*, 51(11):5352 – 5367. Advances in Mixture Models.

Neville, S. E., Ormerod, J. T., and Wand, M. P. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: Pitfalls and remedies. *Electronic Journal of Statistics*, 8(1):1113–1151.

Nolan, J. (2007). *Stable Distributions: Models for Heavy-Tailed Data*. Birkhauser, Boston.

Nott, D. J., Tan, S., Villani, M., and Kohn, R. (2012). Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 21:797–820.

Ormerod, J. T. and Wand, M. P. (2012). Gaussian variational approximate inference for generalized linear mixed models. *J. Comput. Graph. Stat.*, 21:2–17.

Owen, A. (1997). Monte carlo variance of scrambled net quadrature. *SIAM J. Numer. Anal.*, 34:1884–1910.

Paisley, J., Blei, D., and Jordan, M. (2012). Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, Edinburgh, Scotland, UK.

Peters, G., Sisson, S., and Fan, Y. (2012). Likelihood-free Bayesian inference for α-stable models. *Computational Statistics & Data Analysis*, 56(11):3743 – 3756.

Pitt, M. K., Silva, R. S., Giordani, P., and Kohn, R. (2012). On some properties of Markov chain Monte Carlo simulation methods based on the particle filter. *Journal of Econometrics*, 171(2):134–151.

Ranganath, R., Gerrish, S., and Blei, D. M. (2014). Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, volume 33, Reykjavik, Iceland.

Sacks, J. (1958). Asymptotic distribution of stochastic approximation procedures. *The Annals of Mathematical Statistics*, 29(2):373–405.

Salimans, T. and Knowles, D. A. (2013). Fixed-form variational posterior approximation through stochastic linear regression. *Bayesian Analysis*, 8(4):741–908.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84:653–667.

Tan, L. S. L. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28:168–188.

Tavare, S., Balding, D. J., Griffiths, R. C., and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2):505–518.

Tran, M.-N., Kohn, R., Quiroz, M., and Villani, M. (2016). Block-wise pseudo-marginal metropolis-hastings. Technical report. arXiv:1603.02485.

Tran, M.-N., Pitt, M. K., and Kohn, R. (2014). Annealed important sampling for models with latent variables. http://arxiv.org/abs/1402.6035.

Tran, M.-N., Scharth, M., Pitt, M. K., and Kohn, R. (2013). Importance sampling squared for Bayesian inference in latent variable models. http://arxiv.org/abs/1309.3339.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*. Cambridge University Press.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, 15:1351–1369.