JOURNAL OF COMPUTATIONAL AND GRAPHICAL STATISTICS
2018, VOL. 27, NO. 3, 465–478
https://doi.org/10.1080/10618600.2017.1390472

Taylor & Francis
Taylor & Francis Group

Gaussian Variational Approximation With a Factor Covariance Structure

Victor M.-H. Ong[a], David J. Nott[a,b], and Michael S. Smith[c]

[a]Department of Statistics and Applied Probability, National University of Singapore, Singapore; [b]Institute of Operations Research and Analytics, National University of Singapore, Singapore; [c]Melbourne Business School, University of Melbourne, Carlton, VIC, Australia

ABSTRACT

Variational approximations have the potential to scale Bayesian computations to large datasets and highly parameterized models. Gaussian approximations are popular, but can be computationally burdensome when an unrestricted covariance matrix is employed and the dimension of the model parameter is high. To circumvent this problem, we consider a factor covariance structure as a parsimonious representation. General stochastic gradient ascent methods are described for efficient implementation, with gradient estimates obtained using the so-called "reparameterization trick." The end result is a flexible and efficient approach to high-dimensional Gaussian variational approximation. We illustrate using robust P-spline regression and logistic regression models. For the latter, we consider eight real datasets, including datasets with many more covariates than observations, and another with mixed effects. In all cases, our variational method provides fast and accurate estimates. Supplementary material for this article is available online.

ARTICLE HISTORY
Received January 2017
Revised August 2017

KEYWORDS
Gaussian variational approximation; P-spline; Stochastic gradient ascent; variational Bayes

1. Introduction

Variational approximation methods are a promising approach to scalable approximate Bayesian inference in the case of large datasets and highly parameterized models. However, if the variational approximation takes the form of a multivariate Gaussian distribution with an unrestricted covariance matrix, it is difficult to perform variational inference for a high-dimensional parameter because the number of elements in the covariance matrix increases quadratically with the parameter dimension. Hence, in the context of Gaussian variational approximation, it is important to find parsimonious but flexible ways of parameterizing the covariance matrix. The contribution of the present article is to develop general methods for Gaussian variational approximation when the covariance matrix has a factor structure. By general here we mean that the methods do not require any conjugate structure for the model, although it is required that the prior and likelihood and their derivatives are computable. A key feature of our approach is that we obtain efficient gradient estimates for a stochastic gradient ascent optimization procedure using the so-called "reparameterization trick." This leads to a flexible and computationally attractive approach to high-dimensional Gaussian variational approximation.

Let θ be a continuous parameter of dimension m, and consider Bayesian inference with a prior density $p(\theta)$ and likelihood $p(y|\theta)$. Write the posterior density as $p(\theta|y)$, and to simplify notation later write $h(\theta) = p(\theta)p(y|\theta)$, so that $p(\theta|y) \propto h(\theta)$. Variational approximation methods (Attias 1999; Jordan et al. 1999; Winn and Bishop 2005; Ormerod and Wand 2010) provide approximate methods for performing Bayesian calculations that require reduced computation compared to exact methods such as Markov chain Monte Carlo (MCMC). In a variational approach, it is assumed that the posterior density can be approximated by a member of some tractable family of approximations, with typical element $q_\lambda(\theta)$ say, where λ are variational parameters to be chosen indexing different members of the family. Writing $p(y) = \int p(\theta)p(y|\theta)d\theta$ for the marginal likelihood of y, the following identity holds, for any $q_\lambda(\theta)$:

$$\log p(y) = \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)d\theta + \mathrm{KL}(q_\lambda(\theta)||p(\theta|y)), \tag{1}$$

where

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

is the Kullback–Leibler divergence from $q_\lambda(\theta)$ to $p(\theta|y)$. Derivation of Equation (1) can be found, for example, in Ormerod and Wand (2010, p. 42). We denote the expectation with respect to $q_\lambda(\theta)$ as $E_q(\cdot)$. Because the Kullback–Leibler divergence is nonnegative, from Equation (1),

$$\mathcal{L}(\lambda) = \int \log \frac{p(\theta)p(y|\theta)}{q_\lambda(\theta)} q_\lambda(\theta)d\theta = E_q(\log h(\theta) - \log q_\lambda(\theta)) \tag{2}$$

is a lower bound on $\log p(y)$, called the variational lower bound. The Kullback–Leibler divergence is a measure of the quality of the approximation of the true posterior by $q_\lambda(\theta)$, and we choose λ so that the approximation is optimal. Because the left-hand

CONTACT David J. Nott *standj@nus.edu.sg* Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546 and Institute of Operations Research and Analytics, National University of Singapore, 21 Lower Kent Ridge Road, Singapore 119077.
Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/JCGS*.
Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JCGS*.

side of (1) does not depend on the variational parameters, minimizing the Kullback–Leibler divergence $\mathrm{KL}(q_\lambda(\theta)||p(\theta|y))$ with respect to λ is equivalent to maximizing $\mathcal{L}(\lambda)$ with respect to λ. Therefore, maximizing $\mathcal{L}(\lambda)$ with respect to λ provides the best approximation to our posterior distribution within the approximating class in the Kullback–Leibler sense. In this article, we will be concerned with the situation where $q_\lambda(\theta)$ is multivariate normal, so without any further restriction the variational parameters λ consist of both the mean vector and distinct elements of the covariance matrix of the normal variational posterior. As mentioned above, a full normal variational approximation is difficult to work with in high dimensions. Assuming a diagonal covariance structure is one possible simplification, but this loses any ability to represent dependence in the posterior distribution.

Various suggestions in the literature exist for parsimonious ways to parameterize covariance matrices in Gaussian variational approximations. Opper and Archambeau (2009) noted that with a Gaussian prior and a factorizing likelihood, the optimal Gaussian variational distribution can be specified in terms of a much reduced set of variational parameters. Challis and Barber (2013) considered posterior distributions which can be expressed as a product of a Gaussian factor and positive potential, and considered banded Cholesky, Chevron Cholesky, and subspace Cholesky approximations. They are also able to prove concavity of the variational lower bound in this setup. Titsias and Lázaro-Gredilla (2014) considered both full and diagonal covariance structures with the covariance matrix parameterized in terms of the Cholesky factor, where stochastic gradient variational Bayes methods are used. Efficient gradient estimates are constructed using the so-called "reparameterization trick" (Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014). Kucukelbir et al. (2017) considered both unrestricted and diagonal covariance matrices, as well as marginal transformations to improve normality, working in an automatic differentiation environment and using similar gradient estimates to Titsias and Lázaro-Gredilla (2014). Salimans and Knowles (2013) considered a variety of stochastic gradient optimization approaches for learning exponential family type approximations or hierarchical extensions of such approximations. In the Gaussian case, they parameterized in terms of the precision matrix, and exploited sparsity of Hessian matrices for the joint model in their computations. As well as their algorithm using the Hessian, they also provided algorithms that require only computation of first-order derivatives. Archer et al. (2016) considered Gaussian variational approximation in the context of smoothing for state space models. They parameterized the variational optimization in terms of a sparse precision matrix, and exploited the way that this leads to a sparse Cholesky factor in random variate generation from their variational posterior distribution. The blocks of the mean vector and nonzero blocks of the precision matrix are parameterized in terms of global parameters that relate them to local data—an example of so-called amortized variational inference—which was also introduced by Kingma and Welling (2014). Tan and Nott (2018) parameterized the variational optimization directly in terms of the Cholesky factor of the precision matrix and imposed sparsity on the Cholesky factor that reflects conditional independence relationships. They showed how the sparsity can be exploited in the computation of gradients with the reparameterization trick.

In the above work, the approximations considered either require some special structure of the model (such as conditional independence, Gaussian priors, or a factorizing likelihood), do not scale well to high dimensions, or are inflexible in the kinds of dependence they can represent accurately. The goal of the present work is to consider a general method for Gaussian variational approximation, where the covariance matrix is parameterized in terms of a factor structure (Bartholomew, Knott, and Moustaki 2011). By assuming a factor structure, the number of variational parameters is reduced considerably when the number of factors is much less than the full dimension of the parameter space. Such a parsimonious approximation has strong potential in many applications. For example, in models with random effects dependence among the high-dimensional vector of random effects can often be explained by their shared dependence on just a small number of global parameters. We demonstrate this here both for a mixed effects logistic regression model and for a robust P-spline, the latter of which has basis term coefficients that are random effects.

Gaussian variational approximations with a factor covariance structure have been considered previously by Barber and Bishop (1998) and Seeger (2000). However, these authors only consider models in which the variational lower bound can be evaluated analytically, or using one-dimensional numerical quadrature. In contrast, here we consider approaches to performing the required variational optimization without requiring any special structure for the prior or a factorizing likelihood. In independent work Miller, Foti, and Adams (2016) had recently also suggested the use of a factor parameterizations of the covariance matrix in a Gaussian variational approximation, using stochastic gradient methods and the reparameterization trick for gradient estimation. However, their focus is on building mixture of Gaussian variational approximations using boosting and they do not give expressions for the gradient estimates for the Gaussian factor components or the derivation of such results. Rezende, Mohamed, and Wierstra (2014) considered a one-factor approximation to the precision matrix in conjunction with stochastic variational inference, whereas we consider more than one factor. The methods we present here for reparameterizing the variational distribution are complementary to other techniques that can be used to obtain fast approximate inference in Bayesian models. For example, in models where the prior or likelihood is difficult to compute, it may be beneficial to employ approximations to the model itself. This occurs, for example, in sparse approximations of Gaussian process priors (Titsias 2009). Also it may be important to consider different approaches to the optimization and to take advantage of conjugate structure in the model when it exists, for example, see Khan and Lin (2017) who suggested an approach to combining the flexibility of stochastic gradient methods with the ability to exploit conjugate structure using a stochastic mirror descent algorithm.

In the next section, we briefly introduce the main ideas of stochastic gradient variational Bayes. Section 3 then gives details of our stochastic gradient ascent algorithm for optimization of the variational parameters in a Gaussian approximation with factor covariance structure. Efficient gradient estimation based

on the reparameterization trick is developed, and we show that matrix computations in the gradient calculations can be done efficiently using the Woodbury formula. Section 4 illustrates the advantages of the proposed method by first applying it to estimate a robust P-spline in a simulation study, and then to logistic regression using eight real data examples; Section 5 concludes.

2. Stochastic Gradient Variational Bayes

We note that $\mathcal{L}(\lambda)$ in (2) is defined in terms of an expectation, and when this cannot be evaluated in closed form, it is popular to maximize $\mathcal{L}(\lambda)$ using stochastic gradient ascent methods (Ji, Shen, and West 2010; Paisley, Blei, and Jordan 2012; Nott et al. 2012; Hoffman et al. 2013; Salimans and Knowles 2013; Kingma and Welling 2014; Rezende, Mohamed, and Wierstra 2014; Ranganath, Gerrish, and Blei 2014; Titsias and Lázaro-Gredilla 2015). If $\mathcal{L}(\lambda)$ is the objective function to be optimized, $\nabla_\lambda \mathcal{L}(\lambda)$ is its gradient, and $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ is an unbiased estimate of the gradient, then stochastic gradient ascent optimization proceeds as follows. After choosing an initial value $\lambda^{(0)}$ for the variational parameters λ, for $t = 0, 1, \ldots$ perform the update

$$\lambda^{(t+1)} = \lambda^{(t)} + \rho_t \circ \widehat{\nabla_\lambda \mathcal{L}(\lambda^{(t)})},$$

where $\circ$ denotes the Hadamard (element-by-element) product of two random vectors. The update is continued until a stopping condition is satisfied. Here, $\{\rho^{(t)}\}_{t\geq 0}$, is a sequence of vector-valued learning rates, typically chosen so that its elements $\rho_i^{(t)}$ satisfy the Robbins–Monro conditions $\sum_t \rho_i^{(t)} = \infty$ and $\sum_t (\rho_i^{(t)})^2 < \infty$ (Robbins and Monro 1951). Convergence of the sequence $\lambda^{(t)}$ will be to a local optimum under regularity conditions (Bottou 2010). In practice, it is important to consider adaptive learning rates, and in our later examples we implement the ADADELTA approach (Zeiler 2012), although there is a large literature on different adaptive choices of the learning rates.

The references given above differ in the way that the unbiased gradient estimates $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ are constructed, and the variance reduction methods employed. Reducing the variance of the gradient estimates is important because this affects the stability and speed of convergence of the algorithm. Differentiating directly under the integral sign in (2) and using the fact that $E_q(\nabla_\lambda \log q_\lambda(\theta)) = 0$ (the so-called log-derivative trick) and some simple algebra, the gradient is

$$\nabla_\lambda \mathcal{L}(\lambda) = E_q(\nabla_\lambda \log q_\lambda(\theta)(\log h(\theta) - \log q_\lambda(\theta))). \quad (3)$$

Because the expectation is with respect to $q_\lambda(\theta)$, it is easy to estimate (3) unbiasedly using samples from $q_\lambda(\theta)$, provided that sampling from $q_\lambda(\theta)$ is possible. In large datasets, this can also be combined with unbiased estimation of $\log h(\theta)$ using subsampling of terms in the log-likelihood (so-called doubly stochastic variational inference, see Salimans and Knowles 2013; Kingma and Welling 2014; Titsias and Lázaro-Gredilla 2014, e.g.).

In practice, even with sophisticated variance reductions it is often found that derivatives obtained from (3) can have high variance. An alternative approach was considered by Kingma and Welling (2014) and Rezende, Mohamed, and Wierstra (2014), which is usually called the reparameterization trick. To apply this approach, we need to be able to represent samples from $q_\lambda(\theta)$ as $\theta = t(\zeta, \lambda)$, where ζ is a random vector with a fixed density $f(\zeta)$ that does not depend on the variational parameters. In particular, in the case of a Gaussian variational distribution parameterized in terms of a mean vector μ and the Cholesky factor C of its covariance matrix, we can write $\theta = \mu + C\zeta$, where $\zeta \sim N(0, I)$. Then

$$\begin{aligned}\mathcal{L}(\lambda) &= E_q(\log h(\theta) - \log q_\lambda(\theta)) \\ &= E_f(\log h(t(\zeta, \lambda)) - \log q_\lambda(t(\zeta, \lambda))), \end{aligned} \quad (4)$$

where we have written E_f to denote expectation with respect to f. Differentiating under the integral sign in (4) gives an expectation with respect to f that can be estimated unbiasedly based on samples from f. Because of the reparameterization in terms of ζ, the variational parameters have been moved inside the function h so that when we differentiate (4) we are using derivative information from the target posterior density. In practice, it is found that when the reparameterization trick can be applied, it helps greatly to reduce the variance of gradient estimates.

Roeder, Wu, and Duvenaud (2017), generalizing a method described in Han et al. (2016) and Tan and Nott (2018), considered a further approach to variance reduction for gradient estimates. Differentiating under the integral sign in (4) gives

$$\begin{aligned}\nabla_\lambda \mathcal{L}(\lambda) &= E_f\left(\frac{dt(\zeta, \lambda)}{d\lambda}^\top \nabla_\theta \left\{\log h(t(\zeta, \lambda)) - \log q_\lambda(t(\zeta, \lambda))\right\}\right. \\ &\qquad \left. - \nabla_\lambda \log q_\lambda(t(\zeta, \lambda))\right) \quad (5) \\ &= E_f\left(\frac{dt(\zeta, \lambda)}{d\lambda}^\top \nabla_\theta \left\{\log h(t(\zeta, \lambda)) - \log q_\lambda(t(\zeta, \lambda))\right\}\right). \quad (6)\end{aligned}$$

In this expression, if $q_\lambda(\theta)$ is exact so that $h(\theta) \propto q_\lambda(\theta)$, a Monte Carlo estimate of the gradient based on (6) is exactly zero even for one sample, whereas this is not the case for an estimate based on (5). This suggests that, if the variational posterior is a good approximation, then gradient estimation based on (6) is preferred. If the variational family is poorly chosen, or we are far from the mode, then the approximation may not be a good one, and then the additional term in (5) may be considered as acting like a control variate, with an optimal scaling that can be estimated. However, as Roeder, Wu, and Duvenaud (2017) note, this estimation can lead to additional variability, and the simple estimator based on (6) will be considered later. In the lower bound (2), if the variational posterior distribution is Gaussian, we can compute $E_q(\log q_\lambda(\theta))$ analytically, but doing this does not lead to more efficient gradient estimation in general. See Roeder, Wu, and Duvenaud (2017) for further discussion.

3. Approximation with Factor Covariance Structure

In our factor parameterization of the variational distribution, it is assumed that $q_\lambda(\theta) = \phi(\theta; \mu, BB^T + D^2)$ where μ is the mean vector, B is an $m \times p$ full rank matrix with $p << m$, D is a diagonal matrix with diagonal elements $d = (d_1, \ldots, d_m)$, and $\phi(\theta; m, V)$ is an $N(m, V)$ density evaluated at point θ. Without further restrictions B is unidentified, and here we impose the restriction that the upper triangle of B is zero, similar to Geweke and Zhou (1996). For uniqueness we may also wish to impose the restriction on the leading diagonal elements $B_{ii} > 0$,

but we choose not to do so because it does not pose any problem for the variational optimization and it is more convenient to work with the unconstrained parameterization. Note that we can draw $\theta \sim N(\mu, BB^T + D^2)$ by first drawing $\zeta = (z, \epsilon) \sim N(0, I)$ (where z has dimension p, and ϵ has dimension m) and then calculating $\theta = \mu + Bz + d \circ \epsilon$. This will be the basis for our application of the reparameterization trick, and also makes explicit the intuitive idea behind factor models, which is that correlation among the components may be explained in terms of a smaller number of latent variables (z in this case) which influence all the components, with component specific "idiosyncratic" variance being captured through the additional independent error term $d \circ \epsilon$.

We now explain how to apply the reparameterization trick of Kingma and Welling (2014) and Rezende, Mohamed, and Wierstra (2014) to obtain efficient gradient estimates for stochastic gradient variational inference in this setting. Write $f(\zeta)$ for the $N(0, I)$ density of $\zeta = (z^\top, \epsilon^\top)^\top$ in the generative representation of $q_\lambda(\theta)$ described above. The lower bound is an expectation with respect to $q_\lambda(\theta)$, but applying the reparameterization trick and simplifying as in Roeder, Wu, and Duvenaud (2017) gives (6) where $t(\zeta, \lambda) = \mu + Bz + d \circ \epsilon$. Let us write $\mathrm{vec}(B)$ for the vectorization of B obtained by stacking columns and $\otimes$ for the Kronecker product. Using the result $\mathrm{vec}(XYZ) = (Z^T \otimes X)\mathrm{vec}(Y)$ for conformably dimensioned matrices X, Y, and Z, we can write $t(\zeta, \lambda) = \mu + (z^T \otimes I)\mathrm{vec}(B) + d \circ \epsilon$ and then noting that $\nabla_\theta \log q_\lambda(\theta) = -(BB^\top + D^2)^{-1}(\theta - \mu)$,

$$\frac{dt(\zeta, \lambda)}{d\mu} = I \quad \text{and} \quad \frac{dt(\zeta, \lambda)}{d\mathrm{vec}(B)} = z^\top \otimes I, \tag{7}$$

gives

$$\begin{aligned} \nabla_\mu \mathcal{L}(\lambda) &= E_f(\nabla_\theta \log h(\mu + Bz + d \circ \epsilon) \\ &\quad + (BB^\top + D^2)^{-1}(Bz + d \circ \epsilon)), \qquad (8) \\ \nabla_B \mathcal{L}(\lambda) &= E_f(\nabla_\theta \log h(\mu + Bz + d \circ \epsilon) z^T \\ &\quad + (BB^T + D^2)^{-1}(Bz + d \circ \epsilon) z^T). \qquad (9) \end{aligned}$$

Next, noting the symmetry of the way that Bz and $D\epsilon = d \circ \epsilon$ appear in the above expression, we also obtain

$$\begin{aligned} \nabla_d \mathcal{L}(\lambda) = E_f(&\mathrm{diag}(\nabla_\theta \log h(\mu + Bz + d \circ \epsilon)\epsilon^T \\ &+ (BB^T + D^2)^{-1}(Bz + d \circ \epsilon)\epsilon^T)), \end{aligned} \tag{10}$$

where we have written $\mathrm{diag}(Z)$ for the vector of diagonal elements of a square matrix Z. Estimating the expectations in these gradient expressions based on one or more samples from f gives an unbiased estimate $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ of $\nabla_\lambda \mathcal{L}(\lambda)$. This can be used in a stochastic gradient ascent algorithm for optimizing the lower bound, resulting in Algorithm 1.

Initialize $\lambda^{(0)} = (\mu^{(0)}, B^{(0)}, d^{(0)})$, $t = 0$.
Cycle

1. Generate $(\epsilon^{(t)}, z^{(t)}) \sim N(0, I)$
2. Construct unbiased estimates $\widehat{\nabla_\mu \mathcal{L}(\lambda)}$, $\widehat{\nabla_B \mathcal{L}(\lambda)}$, $\widehat{\nabla_d \mathcal{L}(\lambda)}$ of the gradients (8), (9), and (10) at $\lambda^{(t)}$ where the expectations are approximated from the single sample $(\epsilon^{(t)}, z^{(t)})$.
3. Set adaptive learning rate $\rho^{(t)} = (\rho_\mu^{(t)}, \rho_B^{(t)}, \rho_d^{(t)})$ using ADADELTA or other method.
4. Set $\mu^{(t+1)} = \mu^{(t)} + \rho_\mu^{(t)} \circ \widehat{\nabla_\mu \mathcal{L}(\lambda^{(t)})}$.
5. Set $B^{(t+1)} = B^{(t)} + \rho_B^{(t)} \circ \widehat{\nabla_B \mathcal{L}(\lambda^{(t)})}$ for elements of $B^{(t+1)}$ on or below the diagonal, with the upper triangle of $B^{(t+1)}$ fixed at zero.
6. Set $d^{(t+1)} = d^{(t)} + \rho_d^{(t)} \circ \widehat{\nabla_d \mathcal{L}(\lambda^{(t)})}$.
7. Set $\lambda^{(t+1)} = (\mu^{(t+1)}, B^{(t+1)}, d^{(t+1)})$, $t \to t + 1$.

until some stopping rule is satisfied

Algorithm 1: Gaussian variational approximation algorithm with factor covariance structure.

At first sight it may seem that computing the gradient estimates based on (9) and (10) is difficult when the dimension of θ is high because of the inverse of the dense $m \times m$ matrix $(BB^T + D^2)$ in these expressions. However, note that by the Woodbury formula we have

$$(BB^T + D^2)^{-1} = D^{-2} - D^{-2}B(I + B^T D^{-2} B)^{-1} B^T D^{-2}$$

and that on the right-hand side the matrix $(I + B^T D^{-2} B)$ is $p \times p$ with $p << m$ and D is diagonal. So any computation involving $(BB^T + D^2)^{-1}$ or solutions of linear systems in $(BB^T + D^2)$ can be done efficiently in terms of both memory and computation time.

4. Examples

We now demonstrate the advantages of our proposed method, which we call variational approximation with factor covariance structure (VAFC), for both simulated and real examples. In all the examples, we set step sizes (learning rates) adaptively using the ADADELTA method (Zeiler 2012), with different step sizes for each element of λ. Specifically, at iteration $t + 1$, the ith element λ_i of λ is updated as

$$\lambda_i^{(t+1)} = \lambda_i^{(t)} + \Delta\lambda_i^{(t)}.$$

Here, the step size $\Delta\lambda_i^{(t)}$ is $\rho_i^{(t)} g_{\lambda i}^{(t)}$ where $g_{\lambda i}^{(t)}$ denotes the ith component of $\widehat{\nabla_\lambda \mathcal{L}(\lambda^{(t)})}$ and $\rho_i^{(t)}$ is

$$\rho_i^{(t)} = \frac{\sqrt{E\left(\Delta_{\lambda i}^2\right)^{(t-1)} + \epsilon}}{\sqrt{E\left(g_{\lambda i}^2\right)^{(t)} + \epsilon}},$$

where ϵ is a small positive constant, with $E(\Delta_{\lambda i}^2)^{(t)}$ and $E(g_{\lambda i}^2)^{(t)}$ being decayed running average estimates of $\Delta\lambda_i^{(t)^2}$ and $g_{\lambda i}^{(t)^2}$, defined by

$$\begin{aligned} E(\Delta_{\lambda i}^2)^{(t)} &= \upsilon E\left(\Delta_{\lambda i}^2\right)^{(t-1)} + (1 - \upsilon)\Delta\lambda_i^{(t)^2} \\ E\left(g_{\lambda i}^2\right)^{(t)} &= \upsilon E\left(g_{\lambda i}^2\right)^{(t-1)} + (1 - \upsilon) g_{\lambda i}^{(t)^2}. \end{aligned}$$

The variable υ is a decay constant. In the examples we use the default tuning parameter choices $\epsilon = 10^{-6}$ and $\upsilon = 0.95$, and initialize $E(\Delta_{\lambda i}^2)^{(0)} = E(g_{\lambda i}^2)^{(0)} = 0$.

4.1. Simulation Study: Robust P-Spline

In our first example, we consider a P-spline (Eilers and Marx 1996) with a normal mixture error distribution to account for outliers (Smith and Kohn 1996). Consider the nonparametric regression model

$$y_i = g(x_i) + \epsilon_i, \tag{11}$$

where x_i is a scalar covariate, y_i is a scalar response, ϵ_i is a random error, and g is a smooth function that requires estimation. A P-spline approximates g as the linear combination $g(x) = b(x)^T\beta$, where $b(x)$ is a vector of B b-spline basis terms constructed on a grid and evaluated at x, and β are unknown coefficients. We consider $B = 30$ below. To produce a smooth function, the coefficients are regularized using the prior

$$\beta|\tau^2, \psi \sim N(0, (\tau^2)^{-1}P(\psi)^{-1}),$$

where $P(\psi)$ is the band 1 (full rank) precision matrix from an AR(1) model with ψ autoregressive coefficient, and both τ^2 and ψ are hyperparameters. We employ this prior, rather than the random walk prior as in Lang and Brezger (2004), because it ensures a proper distribution for $p(y)$. We assume the error follows a normal mixture model with density $p(\epsilon_i) = 0.95\phi(\epsilon_i; 0, \sigma^2) + 0.05\phi(\epsilon_i; 0, 100\sigma^2)$. This distribution has heavy tails and is able to account for a small proportion of outliers, so that inference on g is robust. To fully specify the model, we assume the priors $\sigma^2 \sim \text{IG}(0.01, 0.01)$, and

$$p(\tau^2, \psi) = p(\tau^2) \times (1/0.99)I(0 < \psi < 0.99),$$

where $p(1/\tau^2)$ is the density of a Weibull distribution with shape parameter 0.5 and scale parameter 2.5, as in Klein and Kneib (2016).

The standard approach to estimate this model is to introduce indicator variables $\delta_i \in \{0, 1\}$ for $i = 1, \ldots, n$, such that $\epsilon_i|\delta_i = 0 \sim N(0, \sigma^2)$ and $\epsilon_i|\delta_i = 1 \sim N(0, 100\sigma^2)$ with priors $p(\delta_i) = 0.95^{(1-\delta_i)}0.05^{\delta_i}$ (Smith and Kohn 1996). Estimation is then undertaken using MCMC and data augmentation for $(\sigma^2, \delta, \beta, \tau^2, \psi)$, where $\delta = (\delta_1, \ldots, \delta_n)$, as outlined in Algorithm 2. We do not give further details of the derivation of MCMC steps as the calculations are standard ones.

1. Sample $\sigma^2|\beta, \delta, y$ from its inverse gamma full conditional distribution,

$$\text{IG}\left(0.01 + \frac{n}{2}, 0.01 + \frac{1}{2}(y^* - X^*\beta)^T(y^* - X^*\beta)\right),$$

where $y^* = D(\delta)^{-1/2}y$, $X^* = D(\delta)^{-1/2}X$, $D(\delta) = \text{diag}(100^{\delta_1}, \ldots, 100^{\delta_n})$ and $X = [b(x_1)|\cdots|b(x_n)]$.

2. For $i = 1, \ldots, n$, sample δ_i from its full conditional distribution in the posterior with β integrated out,

$$p(\delta_i|\delta_{-i}, \sigma^2, y) \propto \left(\frac{1}{10}\right)^{\delta_i} |\Omega_\beta|^{-1/2} \times \exp\left(-\frac{1}{2\sigma^2}(y^{*T}y^* - y^{*T}X^*\Omega_\beta^{-1}X^{*T}y^*)\right) p(\delta_i),$$

where δ_{-i} indicates all components of δ except δ_i and $\Omega_\beta = X^{*T}X^* + \tau^2\sigma^2P(\psi)$.

3. Sample β from its full conditional distribution,

$$\beta|\tau^2, \psi, \sigma^2, \delta, y \sim N(\mu_\beta, \Sigma_\beta),$$

with $\Sigma_\beta = \sigma^2\Omega_\beta^{-1}$, $\mu_\beta = \Omega_\beta^{-1}X^{*T}y^*$.

4. Sample τ^2, ψ using a Metropolis–Hastings step based on a normal approximation to the conditional obtaining the mode using Newton–Raphson optimization.

Algorithm 2: Algorithm for one sweep of a data augmentation MCMC sampling scheme for estimating the posterior of the robust P-spline model.

For large datasets, data augmentation can be slow because the n indicators δ have to be updated at each sweep of the MCMC scheme. Therefore, we consider the VAFC approach to approximating the posterior of $(\beta, \tilde{\sigma}^2, \tilde{\tau}^2, \tilde{\psi})$, with δ integrated out analytically, and, where $\tilde{\sigma}^2 = \log\sigma^2$, $\tilde{\tau}^2 = \log\tau^2$, and $\tilde{\psi} = \log\frac{\psi/0.99}{1-\psi/0.99}$. These transformations are to make the ranges of the parameters unrestricted so that a normal approximation is appropriate in the VAFC approach. We consider a simulation study where there are 100 replicates for each of three test functions. The test functions are similar to those considered in Smith and Kohn (1996),

$$g_1(x) = \sin(4\pi x) \quad g_2(x) = 2x - 1$$
$$g_3(x) = 0.25\phi(x; 0.15, 0.05^2) + 0.25\phi(x; 0.6, 0.2^2),$$

where $x_i \sim$Uniform(0, 1). We simulate datasets according to (11), with $\sigma^2 = 1$ and $n = 200$.

In the VAFC approach we consider $p = 0, \ldots, 10$ factors. Results are shown in Figures 1–3 for each test function. The top left panel in each figure shows the data simulated for the first of the 100 replicates in each case, together with the true mean function. Points marked with a cross are outliers. The top middle panel in each figure shows the data points without outliers, and the posterior means of the robust P-spline function estimates obtained using both MCMC and VAFC with $p = 0$. The dashed lines are obtained as plus and minus twice the estimated posterior standard deviation of g. The figures show that, as far as predictive inference is concerned, even with $p = 0$ the VAFC results are already nearly identical with those obtained by MCMC, so that we do not show the fit for larger values of p.

However, parameter inference improves with larger values of p in terms of uncertainty quantification. To illustrate this, the top right panel in each figure shows boxplots (computed over the 100 replicates) of

$$\log\frac{\text{tr}(\text{cov}(\beta|y))}{\text{tr}(\text{cov}^p(\beta|y))},$$

for $p = 0, 4$, and 10 where $\text{cov}^p(\beta|y)$ is the estimated posterior covariance matrix of β obtained from VAFC with p factors, and $\text{cov}(\beta|y)$ is the posterior covariance estimated by MCMC. A value greater than zero for this quantity indicates an overall underestimation of uncertainty in estimating β, whereas a value less than zero indicates overestimation. The lower row of panels in the three figures shows the same quantity for the parameters $\tilde{\sigma}^2$, $\tilde{\tau}^2$, and $\tilde{\psi}$. We can see there is an underestimation of variability for the parameters $\tilde{\tau}^2$ and $\tilde{\psi}$, the reason being that the posterior distributions for these variables are highly skewed so that the normal approximation works poorly. Experimenting with different transformations could improve this, but these prior hyperparameters are not of primary interest for inference

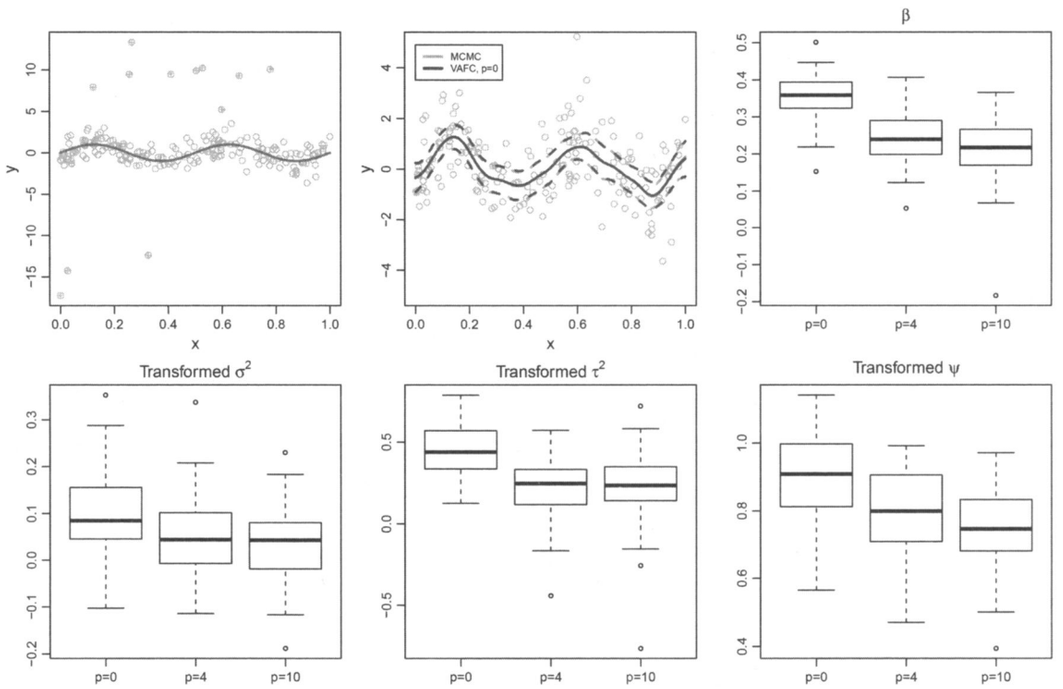

Figure 1. Robust p-spline simulation study results for test function g_1. Top left: Test function g_1, and data from the first simulation study replicate with identified outliers denoted by crosses. Top middle: Data for the first simulation study replicate with outliers excluded, along with posterior mean estimates of g, and bands equal to $g \pm$ two standard deviations of g, estimated using both MCMC and VAFC with $p = 0$. The two estimates are so close, they are hard to discern visually. Top right: Boxplots of the log ratio of the posterior covariance trace for β, for MCMC (numerator) and VAFC (denominator), for $p = 0, 4, 10$. Bottom row: Boxplots of the log ratios of the posterior variances for $\tilde{\sigma}^2$, $\tilde{\tau}^2$, $\tilde{\psi}$ for MCMC (numerator) and VAFC (denominator) for $p = 0, 4, 10$.

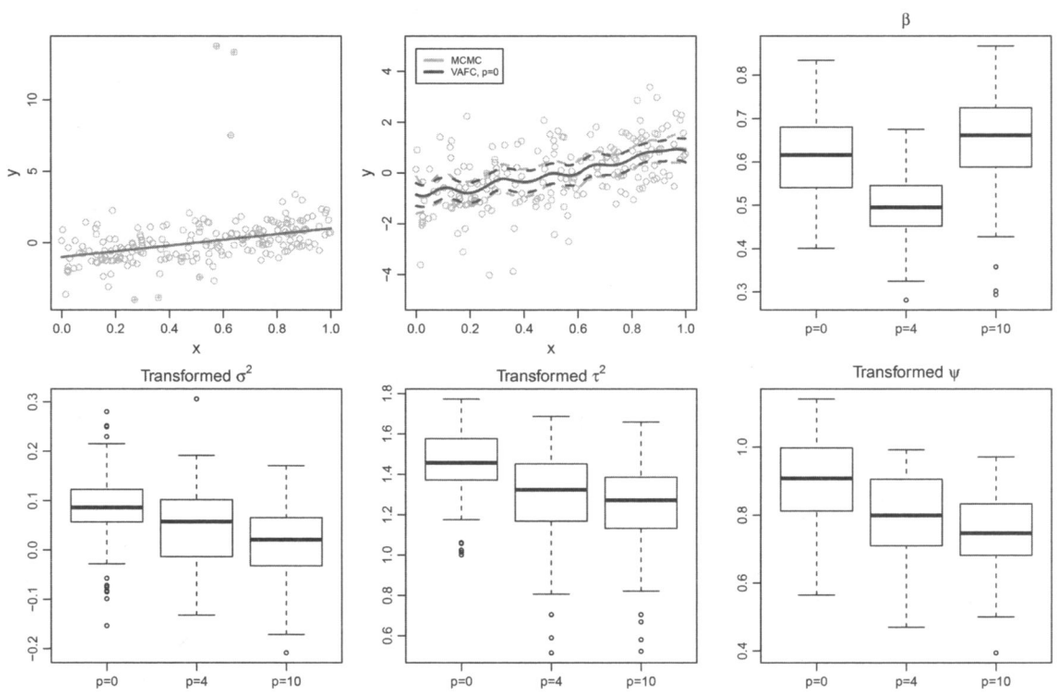

Figure 2. Robust p-spline simulation study results for test function g_2. Top left: Test function g_2, and data from the first simulation study replicate with identified outliers denoted by crosses. Top middle: Data for the first simulation study replicate with outliers excluded, along with posterior mean estimates of g, and bands equal to $g \pm$ two standard deviations of g, estimated using both MCMC and VAFC with $p = 0$. The two estimates are so close, they are hard to discern visually. Top right: Boxplots of the log ratio of the posterior covariance trace for β, for MCMC (numerator) and VAFC (denominator), for $p = 0, 4, 10$. Bottom row: Boxplots of the log ratios of the posterior variances for $\tilde{\sigma}^2$, $\tilde{\tau}^2$, $\tilde{\psi}$ for MCMC (numerator) and VAFC (denominator) for $p = 0, 4, 10$.

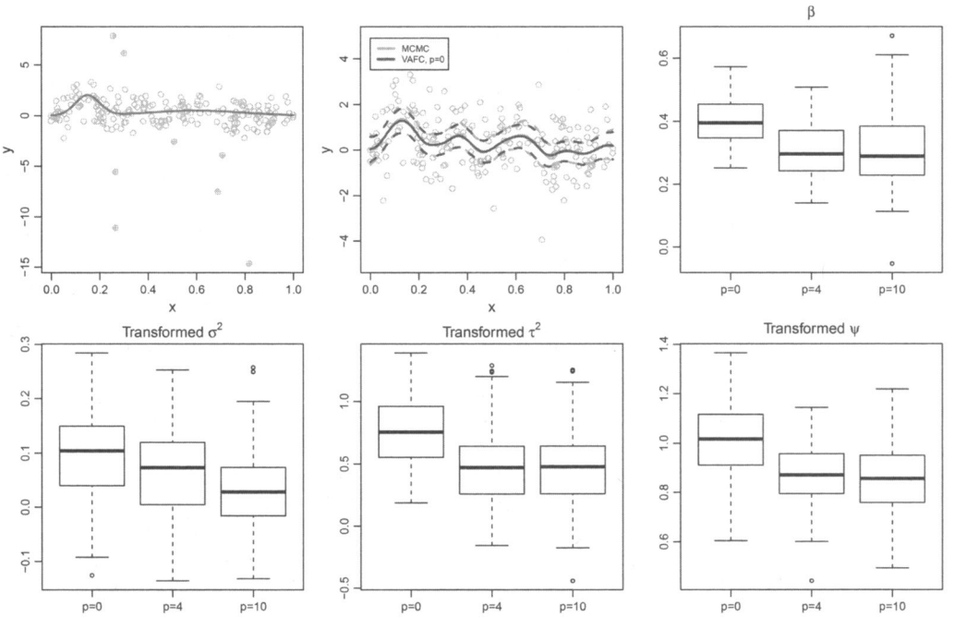

Figure 3. Robust p-spline simulation study results for test function g_3. Top left: Test function g_3, and data from the first simulation study replicate with identified outliers denoted by crosses. Top middle: Data for the first simulation study replicate with outliers excluded, along with posterior mean estimates of g, and bands equal to $g \pm$ two standard deviations of g, estimated using both MCMC and VAFC with $p = 0$. The two estimates are so close, they are hard to discern visually. Top right: Boxplots of the log ratio of the posterior covariance trace for β, for MCMC (numerator) and VAFC (denominator), for $p = 0, 4, 10$. Bottom row: Boxplots of the log ratios of the posterior variances for $\tilde{\sigma}^2$, $\tilde{\tau}^2$, $\tilde{\psi}$ for MCMC (numerator) and VAFC (denominator) for $p = 0, 4, 10$.

and have little effect on predictive inference. For β and σ^2, we see that for $p = 0$ there is generally some underestimation of uncertainty, and this generally improves as we move to $p = 4$ and $p = 10$. The results for $p = 10$ are nearly identical to those obtained using an unrestricted Gaussian approximation (results not shown). Because this is only a relatively low-dimensional example, there is not much difference between the computation times for different values of p in the VAFC method, with all the timings approximately 8 sec for each replicate and test function and 1000 iterations of the stochastic gradient ascent scheme running on a quad processor Windows PC 3.10 GHz workstation. Convergence usually occurs within about 50 iterations, so that 1000 iterations are very conservative. The MCMC scheme, on the other hand, takes about 240 sec for each replicate and test function, for 5000 sampling iterations with 3000 burn in iterations, so the VAFC method provides a roughly 30-fold reduction in computation time.

4.2. Real Data Examples: Logistic Regression

We now consider eight real data examples for a logistic regression model with both a low- and high-dimensional covariate. Our final example (Section 4.2.3) considers logistic regression with random effects for binary longitudinal data. Suppose we are given a dataset with response $y_i \in \{-1, 1\}$ and covariates $\tilde{x}_i \in \mathcal{R}^q$ for $i = 1, \ldots, n$. For a logistic regression, the likelihood is $p(y|\beta) = \prod_{i=1}^n 1/(1 + e^{-y_i x_i^T \beta})$, where $x_i = [1\ \tilde{x}_i^T]^T$ and β denotes the coefficient vector. If θ is a vector containing β and any hyper-parameters, then the posterior is $p(\theta|y) \propto h(\theta) = p(y|\beta)p(\theta)$. Our VAFC approach will be compared with the DSVI (doubly stochastic variational inference) algorithm proposed by Titsias and Lázaro-Gredilla (2014). Similar to VAFC, these authors used a multivariate normal posterior approximation $q^D(\theta) = \phi(\theta; \mu^D, \Sigma^D)$, where the covariance matrix Σ^D is parameterized as $\Sigma^D = C^D C^{D^T}$, with C^D an unrestricted lower triangular Cholesky factor. Both μ^D and C^D can be updated using a stochastic gradient optimization procedure. The VAFC algorithm differs by parameterizing the covariance matrix through a more parsimonious factor structure. We write $q^F(\theta) = \phi(\theta; \mu^F, \Sigma^F)$ for the VAFC posterior approximation.

Four examples in Section 4.2.1 illustrate the performance of DSVI and VAFC when the number of predictors $\tilde{m}$ is moderate and where $\tilde{m} < n$, the kind of situation where there may be most interest in parameter inference and uncertainty quantification. In these examples, we also compare the accuracy of the variational approximations to the exact posterior distribution, computed using MCMC. The three examples in Section 4.2.2 consider cases in which $\tilde{m} >> n$ and where the computational gains from using the factor structure are larger. In these saturated models, we employ a horseshoe prior for parameter shrinkage (Carvalho, Polson, and Scott 2010), so that the variational approximation is to the posterior for the high-dimensional covariate coefficients augmented with the matching local shrinkage parameters. In these examples, interest mostly focuses on predictive inference. Last, in Section 4.2.3 we consider an example for a mixed effects logistic regression model. In this case, the variational approximations are to the posterior augmented with a high-dimensional vector of random effect terms.

4.2.1. Low-Dimensional Logistic Regression

We consider the `spam`, `krkp`, `ionosphere`, and `mushroom` data from the UCI Machine Learning Repository (Lichman 2013). Following Gelman et al. (2008), we change the covariate matrix into binary variables using the `discretization`

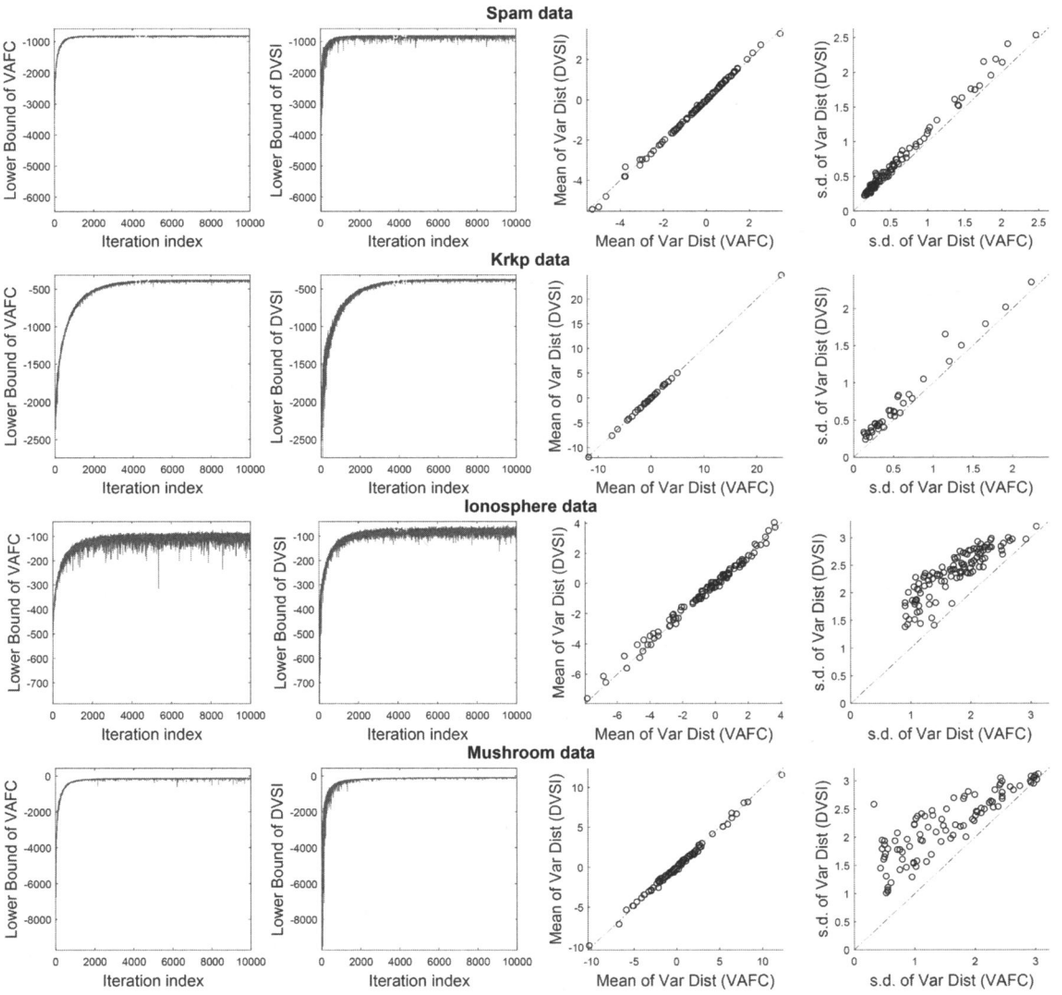

Figure 4. Comparison of VAFC with $p = 3$ factors and the DSVI for four datasets. Each row gives results for a different dataset. The columns give Monte Carlo estimates of the lower bound for VAFC (first column) and DSVI (second column), and a comparison of variational means for VAFC and DSVI (third column) and standard deviations of VAFC and DSVI (fourth column). Points near the (red) 45-degree line in the third column indicate that the variational means are similar for both VAFC and DSVI. Points above the (red) 45 lines in the last column indicate that the variational standard deviations of the regression coefficients are smaller for VAFC than for DSVI.

function (Kim 2016) in R (R Core Team 2017). After doing this, the `spam`, `krkp`, `ionosphere`, and `mushroom` data, respectively, contain $n = 4601, 351, 3196$, and 8124 samples and $\tilde{m} = 104, 111, 37$, and 95 variables, so that $\tilde{m} < n$ in each case. In the examples in this section we use an $N(0, 10I)$ prior for θ.

The first and second columns of Figure 4 show, respectively, Monte Carlo estimates of the lower bounds for DSVI and VAFC with $p = 3$ factors over 10,000 iterations. Convergence is slightly faster for the VAFC method in these examples, and each iteration of the optimization also requires less computation—advantages that are even more pronounced in the high-dimensional case considered in Section 4.2.2. To examine the quality of marginal inferences, the third column of Figure 4 contains a scatterplot of (μ_i^F, μ_i^D) for $i = 1, \ldots, m$ (i.e., the variational means for the two methods) and we see that the values are close. The rightmost column of Figure 4 shows a similar graphical comparison of the estimated posterior standard deviations of the coefficients for VAFC and DSVI, plotting $(\sqrt{\Sigma_{i,i}^F}, \sqrt{\Sigma_{i,i}^D})$ for $i = 1, \ldots, \tilde{m}$. A variational approximation using an insufficiently flexible approximating family often leads to underestimation of posterior variances, for example, see Wang and Titterington (2005). This is indicated here for the VAFC method, with many points appearing above the diagonal lines in the plots. However, this underestimation of the posterior standard deviations is minor, except for the `ionosphere` and `mushroom` datasets. Figure 5 shows what happens when the number of factors in the VAFC method is increased to $p = 20$ for these datasets and, as expected, this reduces the underestimation of the standard deviations in the variational posterior. Although we compare our VAFC method to DSVI in these plots, the DSVI-based inferences are very similar to those for the exact posterior computed using MCMC. This is illustrated in Figure 6 where variational posterior means and standard deviations for DSVI are plotted against posterior means and standard deviations computed using MCMC. For the MCMC computations, we used the package `rstanarm` (Stan Development Team 2016).

In this example, we have considered results of the VAFC method using $p = 3$ and $p = 20$ factors. A reasonable question is how to choose the number of factors in the approximation.

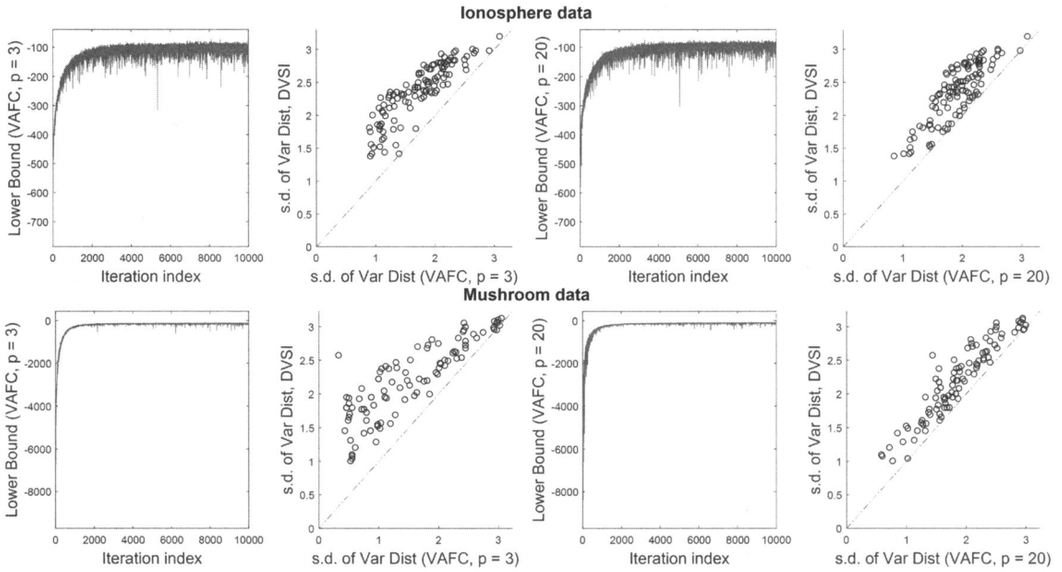

Figure 5. Comparison of DSVI and VAFC with $p = 3$ and $p = 20$ factors for two datasets. Panels in the first and second rows are for the ionosphere and mushroom datasets, respectively. The columns give Monte Carlo estimates of the lower bound for VAFC for $p = 3$ (first column) and $p = 20$ (third column) and comparison of variational standard deviations for $p = 3$ (second column) and $p = 20$ (fourth column). In columns two and four, points near the (red) 45-degree lines indicate that the variational standard deviations are similar for both VAFC and DSVI, whereas points above the (red) 45 degree lines indicate that the variational standard deviations of the regression coefficients are smaller for VAFC than for DSVI.

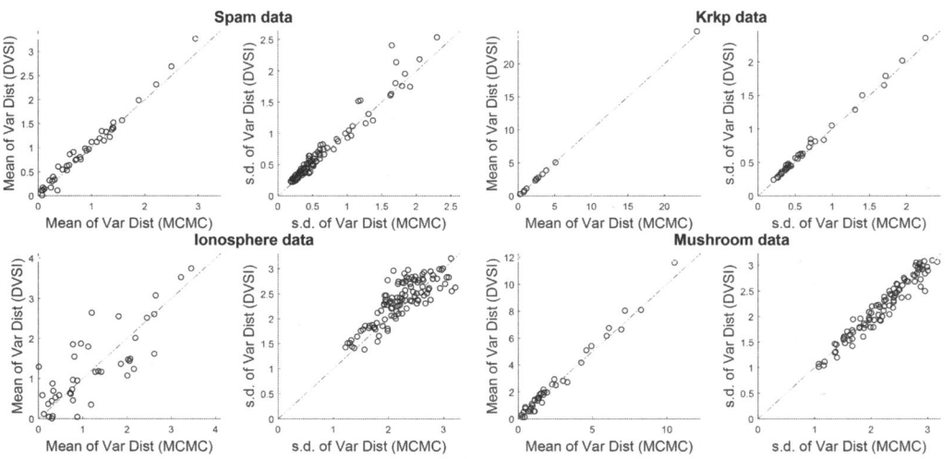

Figure 6. Scatterplots of the points $(\mu_i^{\text{MCMC}}, \mu_i^D)$ and $(\sqrt{\Sigma_{i,i}^{\text{MCMC}}}, \sqrt{\Sigma_{i,i}^D})$ for $i = 1, \ldots, m$ where μ_i^{MCMC} and $\Sigma_{i,i}^{\text{MCMC}}$ uses the MCMC approach `rstanarm`, for the `Spam`, `Krkp`, `Ionosphere`, and `Mushroom` datasets.

One approach is to calculate the approximation for a sequence of increasing values of p, and to stop when posterior inferences of interest no longer change. We consider an approach of this kind further in the example of Section 4.2.3.

Table 1 reports a five-fold cross-validatory assessment of the predictive performance for the four datasets. For the fitted logistic regressions based on μ^D and μ^F, the average training and test set error rates are very similar for the two approaches. This is not surprising given that the variational posterior means tend to be very close for the two methods.

Table 1. Average training and test error rates for the four datasets with $m < n$ estimated via five-fold cross-validation. The VAFC method uses $p = 3$.

	VAFC		DSVI	
Dataset	Training error	Test error	Training error	Test error
Spam	0.046	0.057	0.046	0.057
KRKP	0.027	0.029	0.027	0.030
Ionosphere	0.004	0.082	0.004	0.077
Mushroom data	0	0	0	0

4.2.2. *High-Dimensional Logistic Regression*

We consider the `Colon`, `Leukemia`, and `Breast` cancer datasets available at *http://www.csie.ntu.edu.tw/cjlin/libsvmtools/datasets/binary.html*. The `Colon` dataset has $\tilde{m} = 2000$ covariates with sample sizes of 42 and 20 in the training and test sets, respectively; the `Leukemia` dataset has $\tilde{m} = 7120$ covariates with sample sizes of 38 and 34 in the training and test set; the `Breast` dataset has similar dimension and sample size as the `Leukemia` data in the training set, but with only a sample size of 4 in the test set. The datasets have

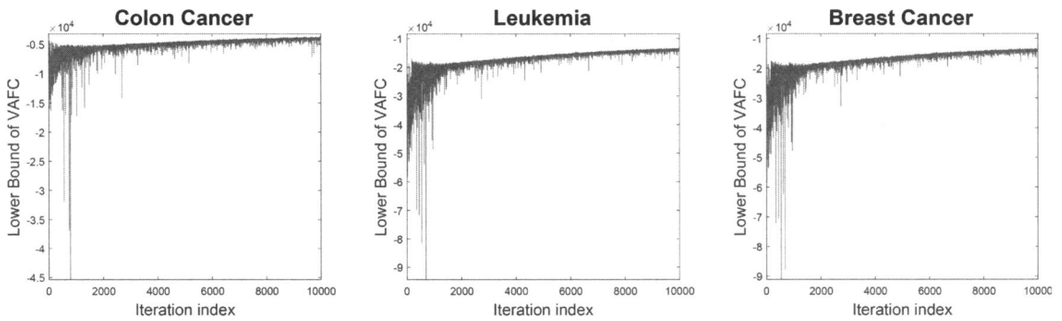

Figure 7. Monte Carlo estimated lower bounds plotted for 10000 SGA iterations for `Colon`, `Leukemia`, and `Breast` cancer datasets using the VAFC method with $p = 4$.

$\tilde{m} >> n$ and the posterior distribution is high dimensional in each case.

Because of the high dimension of the covariate vectors, we consider a sparse signal shrinkage prior distribution on the coefficients, namely, the horseshoe prior (Carvalho, Polson, and Scott 2010). Writing β for the covariate coefficients, we now consider the hierarchical prior

$$\beta_j | g, \delta \sim N(0, \delta_j^2 g^2) \quad \delta_j \sim C^+(0, 1),$$

for $j = 1, \ldots, \tilde{m}$, where $C^+(0, 1)$ denotes the half-Cauchy distribution. The parameters δ_i provide local shrinkage for each coefficient, whereas g is a global shrinkage parameter. For β_0, we use an $N(0, 10)$ prior, and for g we use a half-Cauchy prior, $g \sim C^+(0, 1)$. We let $v = (v_1, \ldots, v_{\tilde{m}+1})^T = (\log \delta_1, \ldots, \log \delta_{\tilde{m}}, \log g)^T$, so that $\theta = (\beta^T, v^T)^T$. We consider a normal variational approximation for θ, using the DSVI and VAFC methods. Mean field variational methods are considered for some applications of the horseshoe and other sparse signal shrinkage priors in Neville, Ormerod, and Wand (2014). Their algorithms do not extend easily to logistic regression, however.

We ran the VAFC algorithm on all three datasets with $p = 0, 4$, and 20. Figure 7 shows a Monte Carlo estimate of the lower bound for the case $p = 4$ versus iteration number for 10,000 iterations. We found that in this example the DSVI algorithm often diverges even with carefully chosen starting values under our prior settings. In terms of computation time, using an iMac computer with i5 3.2 Ghz Intel Quad Core, we found that running 100 iterations of VAFC implemented in `MATLAB` (mat 2012) required approximately 0.12, 0.23, and 0.26 sec for the `Colon`, `Leukaemia`, and `Breast` datasets, respectively, for $p = 0$. Corresponding timings for $p = 4$ were 15.3, 208.2, and 210.0 sec for $p = 4$ and 19.4, 184.5, and 198.8 sec for $p = 20$, respectively. On the other hand, DSVI required 46 sec for the `Colon` data and more than 2 hours for the `Leukaemia` and `Breast` datasets, respectively, for the same number of iterations. The very slow implementation of DSVI for the `Leukaemia` and `Breast` datasets is related to the memory requirements of the DSVI approach, which is another relevant aspect of the comparison of the two algorithms. Note that the timings presented are for the same fixed number of iterations, and the reduced number of variational parameters in the VAFC approach often means than the number of iterations required for convergence is much reduced, so the reduction in computation time is substantial for the VAFC method.

In these high-dimensional examples, Titsias and Lázaro-Gredilla (2014) considered a version of their procedure using a diagonal covariance matrix and a feature selection approach based on automatic relevance determination (DSVI-ARD). We compare predictive performance of the DSVI-ARD approach with the VAFC method with $p = 0$, 4, and 20 factors and the horseshoe prior in Table 2. The DSVI-ARD results are those reported in Titsias and Lázaro-Gredilla (2014). Similar predictive performance is achieved by the two methods. For the leukemia data, out-of-sample predictive performance improves as p increases for the VAFC method. In terms of computation, DSVI-ARD takes about 0.045, 0.060, and 0.071 sec, respectively, for the `Colon`, `Leukaemia`, and `Breast` datasets, respectively, which is slightly faster than VAFC with $p = 0$.

4.2.3. *Mixed Logistic Regression*

In this example, we consider a random intercept model for the polypharmacy dataset described in Hosmer, Lemeshow, and Sturdivant (2013). This longitudinal dataset is available at *http://www.umass.edu/statdata/statdata/stat-logistic.html*, and contains data on 500 subjects, who were followed over 7 years. Following Tan and Nott (2018), we consider a logistic mixed effects model of the form

$$\begin{aligned}&\text{logit } p(y_{ij} = 1|\theta)\\ &\quad = \beta_0 + \beta_{\texttt{gender}}\text{Gender}_i + \beta_{\texttt{race}}\text{Race}_i + \beta_{\texttt{age}}\text{Age}_{ij}\end{aligned}$$

Table 2. Train and test error rates for the three cancer datasets for the VAFC method with $p = 0$, 4, and 20 and the DSVI-ARD method. Errors rates are reported as the ratio of misclassified data points over the number of data points.

	VAFC error rates							
	$p = 0$		$p = 4$		$p = 20$		DSVI-ARD error rates	
Dataset	Training	Test	Training	Test	Training	Test	Training	Test
`Colon`	0/42	0/20	0/42	0/20	0/42	0/20	0/42	1/20
`Leukemia`	0/38	7/34	0/38	6/34	0/38	2/34	0/38	3/34
`Breast`	0/38	1/4	0/38	1/4	0/38	1/4	0/38	2/4

$$+\beta_{M1}\text{MHV1}_{ij}+\beta_{M2}\text{MHV2}_{ij}+\beta_{M3}\text{MHV3}_{ij} \quad (12)$$
$$+\beta_{IM}\text{INPTMHV}_{ij}+u_i$$

for $i = 1, 2, \ldots, 500$ and $j = 1, 2, \ldots, 7$. The response variable y_{ij} is 1 if subject i in year j is taking drugs from three or more different classes, and -1 otherwise. The covariate $\text{Gender}_i = 1$ if subject i is male and 0 if female, while $\text{Race}_i = 0$ if the race of subject i is white and 1 otherwise. Letting MHV_{ij} be the number of outpatient mental health visits for subject i and year j, we construct three dummies: $\text{MHV1}_{ij} = 1$ if $1 \le \text{MHV}_{ij} \le 5$ and 0 otherwise, $\text{MHV2}_{ij} = 1$ if $6 \le \text{MHV}_{ij} \le 14$ and 0 otherwise, and $\text{MHV3}_{ij} = 1$ if $\text{MHV}_{ij} \ge 15$ and 0 otherwise. The covariate INPTMHV_{ij} is 0 if there were no inpatient mental health visits for subject i in year j and 1 otherwise. Finally $u_i \sim N(0, \exp(2\zeta))$ is a subject level random intercept. Write $\beta = (\beta_0, \beta_{\texttt{gender}}, \beta_{\texttt{race}}, \beta_{\texttt{age}}, \beta_{M1}, \beta_{M2}, \beta_{M3}, \beta_{IM})^T$, $u = (u_1, \ldots, u_{500})^T$ and the parameters augmented with the random intercepts as $\theta = (\beta^T, u^T, \zeta)^T$. The prior distribution takes the form

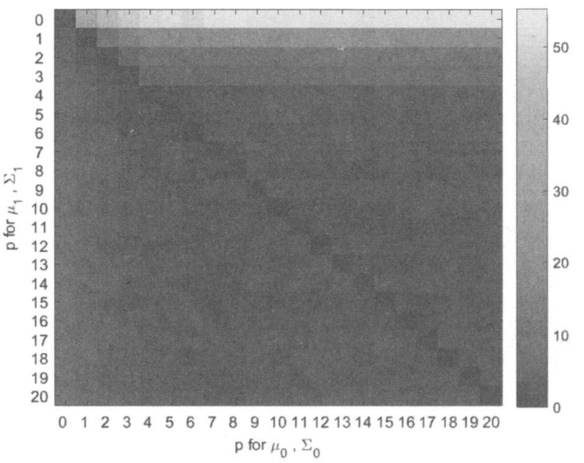

Figure 8. Kullback–Leibler divergences between the final distribution of VAFC for different values of p. The value of p on the x-axis is the one used for taking the expectation in the divergence. The distributions with different values of p are similar for $p \ge 4$.

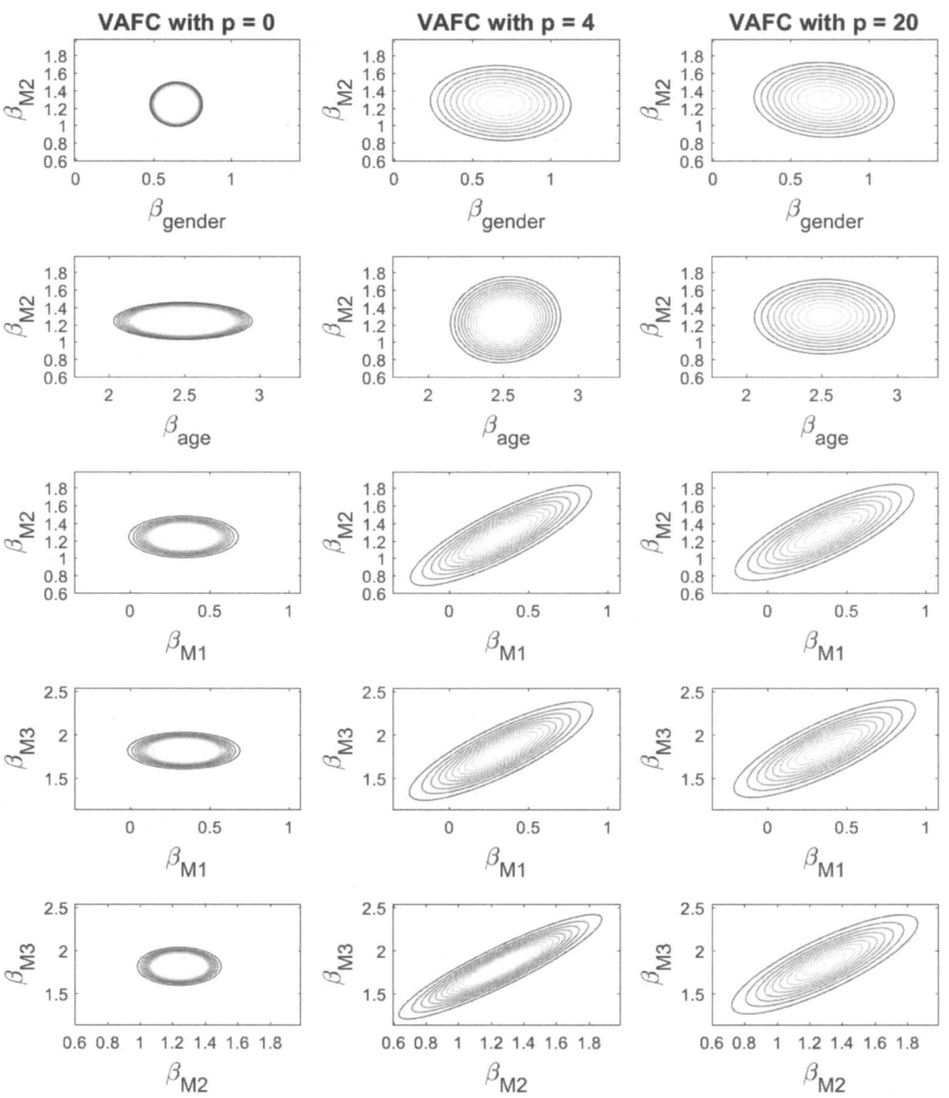

Figure 9. Bivariate contour plots of slices of the variational approximate posterior density for the coefficients with the five highest correlations for the VAFC.

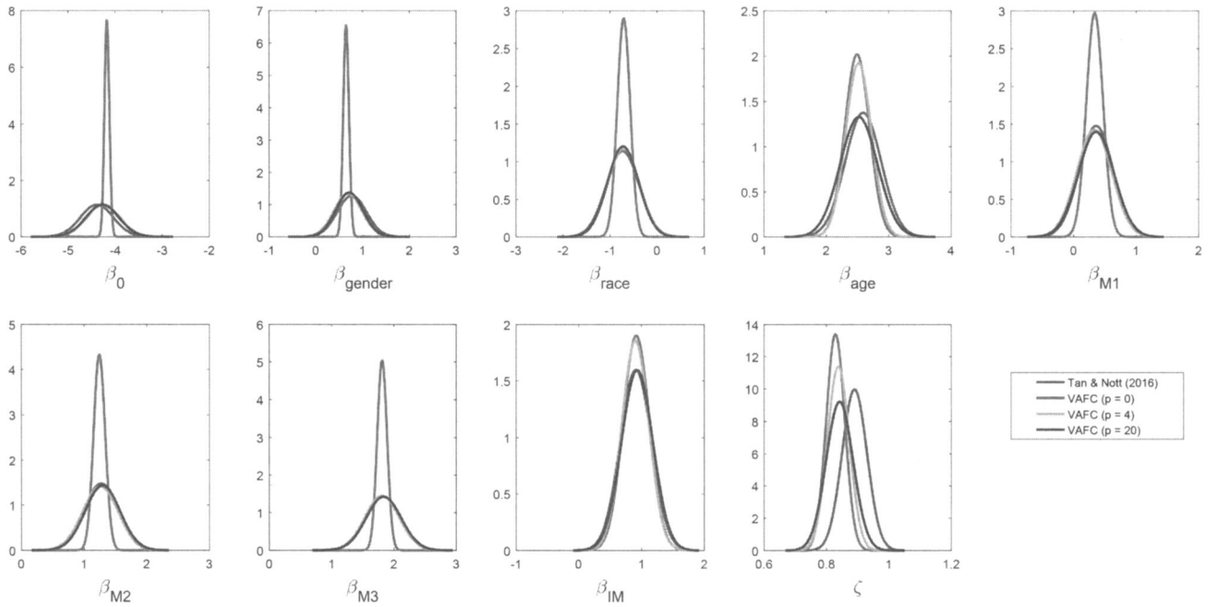

Figure 10. Approximate marginal posterior distributions of components of β and ζ using the method of Tan and Nott (2018) and VAFC with $p = 0$, 4, and 20. Marginal posterior distributions of the regression coefficients are estimated well with $p = 4$.

$$p(\theta) = p(\beta)p(\zeta)\prod_{i=1}^{n} p(u_i|\zeta),$$

where $p(\beta)$ is $N(0, 100I_8)$, $p(\zeta)$ is $N(0, 100)$, and $p(u_i|\zeta)$ is $N(0, \exp(2\zeta))$.

We ran the VAFC algorithm for 10,000 iterations using $p = 0, 1, \ldots, 20$ factors. Figure 8 shows the KL divergences between the variational distributions obtained with different numbers of factors. We can see that the variational distributions are similar for $p \geq 4$. To illustrate this further, Figure 9 shows contour plots of some selected bivariate variational posterior marginals. The results when $p = 0$ (i.e., a diagonal approximation) are very different, and even a crude allowance for posterior correlation with a small number of factors can greatly improve estimation of the posterior marginal distributions. Finally, we also compare the variational marginal density of the regression coefficients with the method in Tan and Nott (2018). The method of Tan and Nott (2018) gives similar answers to MCMC in this example, as shown in Figure 5 of their manuscript, so the Tan and Nott (in press) results can be considered an accurate normal approximation. Figure 10 shows that, except for some mild underestimation of the random intercept variance parameter ζ, the VAFC algorithm with $p = 4$ provides good approximations of the marginal posterior distributions of the components of β—and much better than when $p = 0$. Figure 11 shows plots of the variational posterior means and standard deviations of the subject level random intercepts for VAFC with $p = 4$ against those for the method of Tan and Nott (2018). The posterior distributions of random intercepts are close for the two methods.

5. Discussion

To construct practical variational approximation methods in high dimensions, it is important to employ parsimonious but flexible parameterizations of variational families. Gaussian approximations are important, both because they are useful

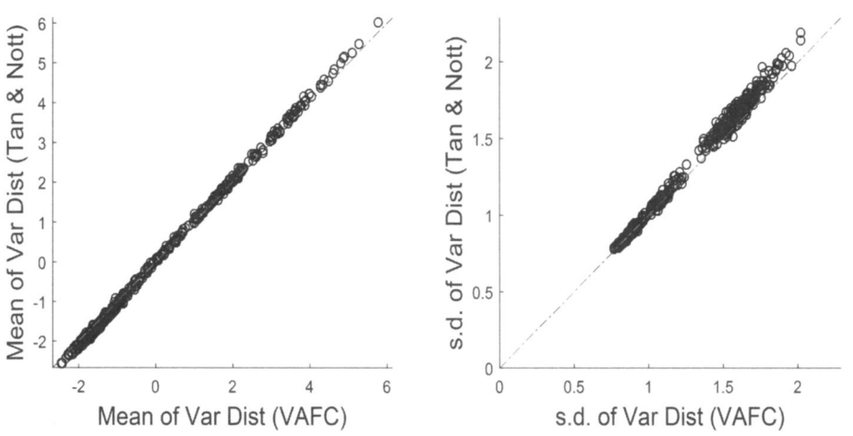

Figure 11. Plot of variational posterior means (left) and standard deviations (right) of the subject level random intercepts for VAFC with $p = 4$ against those for the method of Tan and Nott (2018).

in themselves, but also as a building block for more sophisticated approaches such as variational mixture approximations (Jaakkola and Jordan 1998; Gershman, Hoffman, and Blei 2012; Salimans and Knowles 2013; Guo et al. 2016; Miller, Foti, and Adams 2016) or approximations based on Gaussian copulas (Han et al. 2016). Here, we have considered factor covariance structures for Gaussian variational approximation which may be particularly useful in situations where there is no natural conditional independence structure that can be exploited in the model for reducing the number of free covariance parameters. The approximations can be efficiently formed using the reparameterization trick for gradient estimation and exploiting the Woodbury formula to compute the gradient estimates. In applications to a robust P-spline and a variety of logistic regression models, the methods perform very well.

One difficulty in the application of the presented method relates to the problem of choosing a suitable number of factors. As mentioned in the examples, a useful and obvious heuristic is to apply the method for an increasing sequence of values of p and to stop when inferences of interest no longer change. In applications where a higher level of accuracy is needed, it will be important to go beyond Gaussian approximations of the type considered here, such as using mixture or copula approximations. The recently developed variational boosting approaches of Guo et al. (2016) and Miller, Foti, and Adams (2016) may be particularly useful in this respect. It is also possible in the Gaussian case to combine factor structure with any knowledge of relevant conditional independence relationships in the model. There is room for much ingenuity in exploiting the structure of the model itself for suggesting parsimonious and expressive parameterizations of variational families for particular applications.

Supplementary Materials

MATLAB code (mat 2012) is available online to reproduce the results for the examples.

Acknowledgments

The authors thank Linda Tan for helpful comments on an earlier draft of the article.

Funding

David Nott and Victor Ong were supported by a Singapore Ministry of Education Academic Research Fund Tier 2 grant (R-155-000-143-112).

References

Archer, E., Park, I. M., Buesing, L., Cunningham, J., and Paninski, L. (2016), "Black Box Variational Inference for State Space Models," arXiv:1511.07367. [466]

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*, Morgan Kaufmann, pp. 21–30. [465]

Barber, D., and Bishop, C. M. (1998), "Ensemble Learning for Multi-Layer Networks," in *Advances in Neural Information Processing Systems* (Vol. 10), eds. Jordan, M. I., Kearns, M. J., and Solla, S. A., Cambridge, MA: MIT Press, pp. 395–401 [466]

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011), *Latent Variable Models and Factor Analysis: A Unified Approach* (3rd ed.), New York: Wiley. [466]

Bottou, L. (2010), "Large-Scale Machine Learning with Stochastic Gradient Descent," in *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, Springer, pp. 177–187. [467]

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010), "The Horseshoe Estimator for Sparse Signals," *Biometrika*, 97, 465–480. [471,474]

Challis, E., and Barber, D. (2013), "Gaussian Kullback-Leibler Approximate Inference," *Journal of Machine Learning Research*, 14, 2239–2286. [466]

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing with B-Splines and Penalties," *Statistical Science*, 11, 89–121. [468]

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y.-S. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *The Annals of Applied Statistics*, 2, 1360–1383. [471]

Gershman, S., Hoffman, M. D., and Blei, D. M. (2012), "Nonparametric Variational Inference," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, available at *http://icml.cc/2012/papers/360.pdf* [477]

Geweke, J., and Zhou, G. (1996), "Measuring the Pricing Error of the Arbitrage Pricing Theory," *Review of Financial Studies*, 9, 557–587. [467]

Guo, F., Wang, X., Broderick, T., and Dunson, D. B. (2016), "Boosting Variational Inference," arXiv: 1611.05559. [477]

Han, S., Liao, X., Dunson, D. B., and Carin, L. C. (2016), "Variational Gaussian Copula Inference," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, Cadiz, Spain: JMLR Workshop and Conference Proceedings (Vol. 51), pp. 829–838. [467,477]

Hoffman, M. D., Blei, D. M., Wang, C., and Paisley, J. (2013), "Stochastic Variational Inference," *Journal of Machine Learning Research*, 14, 1303–1347. [467]

Hosmer, D. W., Lemeshow, S., and Sturdivant, R. X. (2013), *Applied Logistic Regression* (3rd ed.), Hoboken, NJ: Wiley. [474]

Jaakkola, T. S., and Jordan, M. I. (1998), "Improving the Mean Field Approximation Via the Use of Mixture Distributions," in *Learning in Graphical Models*, ed. M. I. Jordan, Dordrecht, Netherlands: Springer, pp. 163–173. [477]

Ji, C., Shen, H., and West, M. (2010), "Bounded Approximations for Marginal Likelihoods," Technical Report no. 10-05, Institute of Decision Sciences, Duke University. [467]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [465]

Khan, M. E., and Lin, W. (2017), "Conjugate-Computation Variational Inference: Converting Variational Inference in Non-Conjugate Models to Inferences in Conjugate Models," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ed. A. Singh and J. Zhu, Vol. 54, Proceedings of Machine Learning Research Series, Ft. Lauderdale, FL: PMLR, pp. 878–887. [466]

Kim, H. (2016), "discretization: Data Preprocessing, Discretization for Classification," available at *https://CRAN.R-project.org/package=discretization* [472]

Kingma, D. P., and Welling, M. (2014), "Auto-Encoding Variational Bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR) 2014*, available at *https://arxiv.org/abs/1312.6114* [466,467,468]

Klein, N., and Kneib, T. (2016), "Scale-Dependent Priors for Variance Parameters in Structured Additive Distributional Regression," *Bayesian Analysis*, 11, 1071–1106. [469]

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017), "Automatic Differentiation Variational Inference," *Journal of Machine Learning Research*, 18, 1–45. [466]

Lang, S., and Brezger, A. (2004), "Bayesian P-Splines," *Journal of Computational and Graphical Statistics*, 13, 183–212. [469]

Lichman, M. (2013), *UCI Machine Learning Repository*, Irvine, CA: University of California, School of Information and Computer Sciences. Available at *http://archive.ics.uci.edu/ml* [471]

Miller, A. C., Foti, N. J., and Adams, R. P. (2016), "Variational Boosting: Iteratively Refining Posterior Approximations," in *Proceedings*

of the 34th International Conference on Machine Learning, Sydney, NSW, Australia, 6–11 August, 2420–2429. Available online at *http://proceedings.mlr.press/v70/miller17a.html*. [466,477]

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, 1113–1151. [474]

Nott, D. J., Tan, S. L., Villani, M., and Kohn, R. (2012), "Regression Density Estimation with Variational Methods and Stochastic Approximation," *Journal of Computational and Graphical Statistics*, 21, 797–820. [467]

Opper, M., and Archambeau, C. (2009), "The Variational Gaussian Approximation Revisited," *Neural Computation*, 21, 786–792. [466]

Ormerod, J. T., and Wand, M. P. (2010), "Explaining Variational Approximations," *The American Statistician*, 64, 140–153. [465]

Paisley, J. W., Blei, D. M., and Jordan, M. I. (2012), "Variational Bayesian Inference with Stochastic Search," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, available at *http://icml.cc/2012/papers/687.pdf* [467]

Ranganath, R., Gerrish, S., and Blei, D. M. (2014), "Black Box Variational Inference," in *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics*, JMLR Workshop and Conference Proceedings (Vol. 33), pp. 814–822. [467]

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [472]

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014), "Stochastic Backpropagation and Approximate Inference in Deep Generative Models," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2014*, available at *proceedings.mlr.press/v32/rezende14.pdf* [466,467,468]

Robbins, H., and Monro, S. (1951), "A Stochastic Approximation Method," *The Annals of Mathematical Statistics*, 22, 400–407. [467]

Roeder, G., Wu, Y., and Duvenaud, D. K. (2017), "Sticking the Landing: Simple, Lower-Variance Gradient Estimators for Variational Inference," in *Advances in Neural Information Processing Systems*, Vol. 30, ed. I. Guyo, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Curran Associates, Inc., 6925–6934. Available online at *http://papers.nips.cc/paper/7268-sticking-the-landing-simple-lower-variance-gradient-estimators-for-variational-inference.pdf*. [467,468]

Salimans, T., and Knowles, D. A. (2013), "Fixed-Form Variational Posterior Approximation through Stochastic Linear Regression," *Bayesian Analysis*, 8, 837–882. [466,467,477]

Seeger, M. (2000), "Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers," in *Advances in Neural Information Processing Systems* (Vol. 12), eds. Solla, S. A., Leen, T. K., and Müller, K., Cambridge, MA: MIT Press, pp. 603–609. [466]

Smith, M., and Kohn, R. (1996), "Nonparametric Regression using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343. [468,469]

Stan Development Team (2016), *RStan: the R interface to Stan, Version 2.10.1*, available at *http://mc-stan.org* [472]

Tan, L. S. L., and Nott, D. J. (2018), "Gaussian Variational Approximation with Sparse Precision Matrices," *Statistics and Computing*, *28*, 259–275. [466,467,474,476]

The Mathworks, Inc. (2012), *MATLAB Version 8 (R2012b)*, Natick, MA: The Mathworks, Inc. [474]

Titsias, M. K. (2009), "Variational Learning of Inducing Variables in Sparse Gaussian Processes," in *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, ed. D. van Dyk and M. Welling, Vol. 5, Proceedings of Machine Learning Research Series, Clearwater Beach, FL: PMLR pp. 567–574. [466]

Titsias, M., and Lázaro-Gredilla, M. (2014), "Doubly Stochastic Variational Bayes for Non-Conjugate Inference," in *Proceedings of the 29th International Conference on Machine Learning, ICML 2014*, available at *proceedings.mlr.press/v32/titsias14.pdf* [466,467,471,474]

—— (2015), "Local Expectation Gradients for Black Box Variational Inference," in *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, Curran Associates, Inc., pp. 2638–2646. [467]

Wang, B., and Titterington, D. M. (2005), "Inadequacy of Interval Estimates Corresponding to Variational Bayesian Approximations," in *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics*, Society for Artificial Intelligence and Statistics, pp. 373–380. [472]

Winn, J., and Bishop, C. M. (2005), "Variational Message Passing," *Journal of Machine Learning Research*, 6, 661–694. [465]

Zeiler, M. D. (2012), "ADADELTA: An Adaptive Learning Rate Method," arXiv: 1212.5701. [467,468]