Covariances, Robustness, and Variational Bayes

Ryan Giordano
Department of Statistics
UC Berkeley

Tamara Broderick
Department of EECS
MIT

Michael I. Jordan
Departments of EECS and Statistics
UC Berkeley

September 7, 2017

Abstract

Variational Bayes (VB) is an approximate Bayesian posterior inference technique that is increasingly popular due to its fast runtimes on large-scale datasets. However, even when VB provides accurate posterior means for certain parameters, it often mis-estimates variances and covariances. Furthermore, prior robustness measures have remained undeveloped for VB. By deriving a simple formula for the effect of infinitesimal model perturbations on VB posterior means, we provide both improved covariance estimates and local robustness measures for VB, thus greatly expanding the practical usefulness of VB posterior approximations. The estimates for VB posterior covariances rely on a result from the classical Bayesian robustness literature relating derivatives of posterior expectations to posterior covariances. Our key assumption is that the VB approximation provides good estimates of a select subset of posterior means—an assumption that has been shown to hold in many practical settings. In our experiments, we demonstrate that our methods are simple, general, and fast, providing accurate posterior uncertainty estimates and robustness measures with runtimes that can be an order of magnitude smaller than MCMC.

Key phrases: Variational Bayes; Bayesian robustness; Mean field approximation; Linear response theory; Automatic differentiation

1 Introduction

Most Bayesian posteriors cannot be calculated analytically, so in practice we turn to approximations. Variational Bayes (VB) casts posterior approximation as an optimization problem in which the objective to be minimized is the divergence, among a tractable sub-class of posteriors, from the exact posterior. For example, one widely-used and relatively simple flavor of VB is "mean field variational Bayes" (MFVB), which employs Kullback-Liebler (KL) divergence and a factorizing exponential family approximation for the tractable sub-class of posteriors [Wainwright and Jordan, 2008]. MFVB has been increasingly popular as an alternative to Markov Chain Monte Carlo (MCMC) in part due to its fast runtimes on large-scale data sets. Although MFVB does not come with any general accuracy guarantees (except asymptotic ones in special cases, as in Westling and McCormick [2015], Wang and Blei [2017]), MFVB produces posterior mean estimates of certain parameters that are accurate enough to be useful in a number of real-world applications [Blei et al., 2016]. In the current work we will focus on MFVB for motivation and in our examples and experiments, though many of the ideas here could be applied to more general approximations or divergences.

Despite its computational advantages, MFVB typically underestimates marginal variances [MacKay, 2003, Wang and Titterington, 2004, Turner and Sahani, 2011], and, to our knowledge, techniques for assessing Bayesian robustness have not yet been developed for VB. It is these inferential issues that arethe focus of the current paper.

In contrast to the optimization approach of VB, MCMC constructs a Markov chain with the exact posterior as its stationary distribution. At first glance, it may seem that MCMC is made for integration and MFVB is made for differentiation. Using MCMC draws, integrals with respect to the posterior can be readily approximated by constructing the corresponding sample moment from the MCMC draws. Similarly, as a parametric optimization problem, MFVB lends itself naturally to sensitivity analysis, since its optimum can be differentiated analytically. However, a key result of the Bayesian local robustness literature is that derivatives and covariances are two sides of the same coin, since, under mild regularity conditions, derivatives of posterior quantities can re-cast as posterior covariances by exchanging the order of integration and differentiation (Gustafson [1996b], Basu et al. [1996], Efron [2015] and Section (2.1) below).

Thus, in order to calculate local prior sensitivity, the Bayesian robustness literature re-casts prior sensitivities as posterior covariances that can be easily calculated with MCMC. In order to provide covariance estimates for MFVB, we turn this idea on its head and use the sensitivity of MFVB posterior expectations to estimate their covariances. Additionally, we derive straightforward general VB (not only MFVB) versions of a number of prior sensitivity measures from the Bayesian robustness literature, including parametric sensitivity and sensitivity to additive functional perturbations. Our key assumption is that the posterior means of interest are well-estimated by VB for all the perturbations of interest. In our experiments, we compare MFVB models to both MCMC and maximum a posteriori (MAP) posterior approximations. We find that the MFVB means and variances, unlike the MAP estimates, match the MCMC approximations closely while still running over an order of magnitude faster than MCMC.

We will begin in Section 2 by discussing the general relationship between Bayesian sensitivity and posterior covariance, and then defining local robustness and sensitivity. Next, we will introduce VB and derive the linear system for the VB local sensitivity estimates. In Section 3, we show how to use the VB local sensitivity results to estimate covariances and calculate a range of canonical Bayesian prior sensitivity measures. Finally, in Section 4, we describe our experiments on real industry data that illustrate the speed and effectiveness of our methods. Section 5 concludes.

2 Bayesian (and variational Bayesian) covariances and sensitivity

An MCMC posterior estimate is an empirical distribution formed with posterior draws, and a VB posterior estimate takes the form of a parameterized distribution with optimized parameters. MCMC draws lend themselves naturally to the approximate calculation of posterior moments, such as those required for covariances. In contrast, VB approximations lend themselves naturally to sensitivity analysis, since we can analytically differentiate the optima with respect to perturbations. However, the contrast between derivatives and moments is not so stark since, under mild regularity conditions that allow the exchange of integration and differentiation, there is a direct correspondence between sensitivity and covariance. Thus, as has long been known in the Bayesian robustness literature, we can use the sample covariances of MCMC to calculate posterior sensitivities. With VB approximations, for which sensitivity is more natural, we can apply this relationship in reverse: we can use the VB sensitivity to calculate covariances.

2.1 Covariances and sensitivity

We will first state a general result relating sensitivity and covariance and apply it to our specific cases of interest as they arise throughout the paper. Denote an unknown model parameter by the vector $\theta \in \mathbb{R}^K$, and assume a dominating measure for θ given by λ. Define $\rho(\theta, t)$ to be any λ-measurable function on θ that depends on an index $t \in \mathbb{R}^D$, and assume that $0 < \int \exp(\rho(\theta, t))\,\lambda(d\theta) < \infty$. For our purposes, one may think of $\exp(\rho(\theta, t))$ as the product of a parameterized likelihood and a prior with dependencies other than t left implicit. Later on, we will choose t and $\rho(\theta, t)$ to represent various prior and model perturbations, but for now we will keep the discussion abstract. After normalization we get a density, $p(\theta|t)$, in θ with respect to λ :

$$p(\theta|t) := \frac{\exp(\rho(\theta, t))}{\int \exp(\rho(\theta', t))\,\lambda(d\theta')}.$$

Suppose we are interested in differentiating the expectation

$$\mathbb{E}_{p(\theta|t)}[g(\theta)] := \int p(\theta|t)\, g(\theta)\, \lambda(d\theta)$$

with respect to t at $t = 0$. This will measure the local sensitivity of $\mathbb{E}_{p(\theta|t)}[g(\theta)]$ to the index t at $t = 0$.

Theorem 2.1. *When Assumption (A.1) and Assumption (A.2) hold for all t in a neighborhood of zero, then*

$$\left.\frac{d\mathbb{E}_{p(\theta|t)}[g(\theta)]}{dt^\intercal}\right|_{t=0} = \mathrm{Cov}_{p(\theta|0)}\left(g(\theta), \left.\frac{\partial \rho(\theta, t)}{\partial t}\right|_{t=0}\right). \tag{1}$$

See Appendix A for a proof and technical conditions. By using MCMC draws to calculate the covariance on the right-hand side of Eq. (1), one can form an estimate of $d\mathbb{E}_{p(\theta|t)}[g(\theta)]/dt^\intercal$ at $t = 0$. One might also approach the problem of calculating $d\mathbb{E}_{p(\theta|t)}[g(\theta)]/dt^\intercal$ using importance sampling [Owen, 2013, Chapter 9] as follows. First, an importance sampling estimate of the dependence of $\mathbb{E}_{p(\theta|t)}[g(\theta)]$ on t can be constructed with weights that depend on t. Then, by differentiating the weights with respect to t, this would provide a sample-based estimate of $d\mathbb{E}_{p(\theta|t)}[g(\theta)]/dt^\intercal$. We show in Appendix B that this is approach is equivalent to using MCMC samples to estimate the covariance in Theorem (2.1).

In Section (2.2), Theorem (2.1) will immediately allow us to calculate local sensitivity from MCMC draws using sample covariances. After a little more work, in Section (3.1), Theorem (2.1) will allow us to use the sensitivity of VB means to calculate their covariances.

2.2 Local sensitivity and robustness

As in Section (2.1), denote an unknown model parameter by the vector $\theta \in \mathbb{R}^K$, and assume a dominating measure on θ given by λ. Denote the prior parameters by α, where $\alpha \in \mathbb{R}^M$. Finally, denote our data by x. Suppose we have tentatively chosen a likelihood, $p(x|\theta)$, and a prior, $p(\theta|\alpha)$. We then let p^x_α denote the posterior distribution of θ given x, as given by Bayes' Theorem:

$$p^x_\alpha(\theta) := p(\theta|x, \alpha) = \frac{p(x|\theta)\, p(\theta|\alpha)}{\int p(x|\theta')\, p(\theta'|\alpha)\, \lambda(d\theta')}.$$

We will assume that we are interested in a posterior expectation of some function $g(\theta)$ (e.g., a parameter mean, a posterior predictive value, or squared loss): $\mathbb{E}_{p^x_\alpha}[g(\theta)]$. In the current work, we will quantify the

uncertainty of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ by the posterior variance, $\mathrm{Var}_{p_\alpha^x}(g(\theta))$. Other measures of central tendency (e.g., posterior medians) or uncertainty (e.g., posterior quantiles) may also be good choices, but are beyond the scope of the current work.

Note the dependence of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ on both the likelihood and prior through Bayes' theorem. The choice of a prior and choice of a likelihood is made by the modeler and is almost invariably a simplified representation of the real world. The choice is therefore to some extent subjective, and so one hopes that the salient aspects of the posterior would not vary under reasonable variation in the choice of prior and the choice of likelihood. Consider the prior, for example: the process of prior elicitation may be prohibitively time-consuming; two practitioners may have irreconcilable subjective prior beliefs; or the model may be so complex and high-dimensional that humans cannot reasonably express their prior beliefs as formal distributions. All of these circumstances might give rise to a range of reasonable prior choices. A posterior quantity is "robust" to the prior to the extent that it does not change much when calculated under these different prior choices. Although robustness to the likelihood is no less important, in this paper we will focus on prior robustness, in part for continuity with existing literature.

Quantifying the sensitivity of the posterior to variation in the likelihood and prior is one of the central concerns of the field of robust Bayes [Berger et al., 2000]. (We will not discuss the other central concern, which is the selection of priors and likelihoods that lead to robust estimators.) Suppose that we have determined that the prior parameter α belongs to some set $\mathcal{A}$, perhaps after expert prior elicitation. Ideally, we would calculate the extrema of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ as α ranges over all of $\mathcal{A}$. This is called *global robustness* and is intractable or difficult except in special cases [Moreno, 2000, Huber, 2011, Chapter 15]. An alternative is to examine how much $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ changes locally in response to small perturbations in the value of α. To this end, we define the *local sensitivity* [Gustafson, 2000]:

Definition 2.2. The local sensitivity of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ to prior parameter α is given by

$$\mathbf{S}_\alpha := \left.\frac{d\mathbb{E}_{p_\alpha^x}[g(\theta)]}{d\alpha}\right|_\alpha. \tag{2}$$

$\mathbf{S}_\alpha$, the local sensitivity, can be considered a measure of *local robustness* [Gustafson, 2000]. Throughout the paper we will distinguish between sensitivity, which comprises objectively defined quantities such as $\mathbf{S}_\alpha$, and robustness, which we treat as a more subjective decision that may be informed by the sensitivity as well as other considerations. For example, even if one knows $\mathbf{S}_\alpha$ precisely, how much posterior change is too much change and how much prior variation is reasonable remain decisions to be made by the modeler. For a more in-depth discussion of how we use the terms sensitivity and robustness, see Appendix C.

Local sensitivity can be thought of as quantifying sensitivity to priors within a small region where the posterior dependence on the prior is approximately linear. It provides an approximation to global robustness in the sense that, to first order, for $\Delta\alpha \in \mathcal{A} - \alpha$,

$$\mathbb{E}_{p_\alpha^x}[g(\theta)]\big|_{\alpha+\Delta\alpha} \approx \mathbb{E}_{p_\alpha^x}[g(\theta)] + \mathbf{S}_\alpha^\intercal \Delta\alpha.$$

An immediate corollary of Theorem (2.1) allows us to calculate $\mathbf{S}_\alpha$ as a covariance.

Corollary 2.3. *When the conditions of Theorem (2.1) hold for $t = \alpha$ and $\rho(\theta, \alpha) = \log p(x|\theta) + \log p(\theta|\alpha)$ then*

$$\mathbf{S}_\alpha = \mathrm{Cov}_{p_\alpha^x}\left(g(\theta), \frac{\partial \log p(\theta|\alpha)}{\partial\alpha}\right). \tag{3}$$

See also Basu et al. [1996], in which Corollary (2.3) is stated in the proof of Theorem 1, as well as Pérez et al. [2006] and Efron [2015]. Given MCMC draws from a chain we assume to have reached equilibrium with stationary distribution p_α^x, one can calculate an estimate of $\mathbf{S}_\alpha$ using the sample covariance version of Eq. (1):

$$\hat{\mathbf{S}}_\alpha := \frac{1}{N_s}\sum_{n=1}^{N_s} g(\theta_n)\frac{\partial \log p(\theta_n|\alpha)}{\partial \alpha^\intercal} - \left(\frac{1}{N_s}\sum_{n=1}^{N_s} g(\theta_n)\right)\left(\frac{1}{N_s}\sum_{n=1}^{N_s}\frac{\partial \log p(\theta_n|\alpha)}{\partial \alpha^\intercal}\right) \tag{4}$$

$$\text{for } \theta_n \overset{iid}{\sim} p_\alpha^x(\theta), \text{ for } n = 1, ..., N_s.$$

2.3 Variational Bayes

We now briefly review VB and state our key assumptions about its accuracy. We wish to find an approximate distribution, in some class $\mathcal{Q}$ of tractable distributions, selected to minimize the Kullback-Liebler divergence (KL divergence) between $q \in \mathcal{Q}$ and the exact posterior p_α^x. We assume that distributions in $\mathcal{Q}$ are parameterized by a finite-dimensional parameter η in some feasible set $\Omega_\eta \subseteq \mathbb{R}^{K_\eta}$:

$$\mathcal{Q} \quad := \quad \{q : q = q(\theta;\eta) \text{ for } \eta \in \Omega_\eta\}. \tag{5}$$

Given $\mathcal{Q}$, we define the optimal $q \in \mathcal{Q}$, which we call q_α^x, as that distribution that minimizes the KL divergence $KL(q(\theta;\eta)\,||p_\alpha^x(\theta))$ from p_α^x. We denote the corresponding optimal variational parameters as η^*.

Definition 2.4. The variational approximation $q_\alpha^x(\theta)$ to $p_\alpha^x(\theta)$ is defined by

$$q_\alpha^x(\theta) = q(\theta;\eta^*) \quad := \quad \text{argmin}_{q\in\mathcal{Q}}\{KL(q(\theta;\eta)\,||p_\alpha^x(\theta))\} \tag{6}$$

where

$$KL(q(\theta;\eta)\,||p_\alpha^x(\theta)) = \mathbb{E}_{q(\theta;\eta)}[\log q(\theta;\eta) - \log p(x|\theta) - \log p(\theta|\alpha)] + \log p(x).$$

In the KL divergence, the (generally intractable) normalizing term $\log p(x)$ does not depend on $q(\theta)$ and so can be neglected in the optimization. In order for the KL divergence to be well defined, we assume that both $p(\theta)$ and $q(\theta)$ are given with respect to the same base measure, λ, and that the support of $q(\theta)$ is contained in the support of $p(\theta)$. We additionally assume that $KL(q(\theta;\eta)\,||p_\alpha^x(\theta))$ is twice differentiable as a function of η, that the optimal η^* is interior to Ω_η, and that η^* varies smoothly in α (rigorous statements of these assumptions can be found in Appendix D).

A common choice for the approximating family $\mathcal{Q}$ in Eq. (5) is the "mean field family" [Wainwright and Jordan, 2008, Blei et al., 2016],

$$\mathcal{Q}_{mf} := \left\{q(\theta) : q(\theta) = \prod_k q(\theta_k;\eta_k)\right\}, \tag{7}$$

where k indexes a partition of the full vector θ and of the parameter vector η. That is, $\mathcal{Q}_{mf}$ approximates the posterior p_α^x as a distribution that factorizes across sub-components of θ. Note that, in general, each function $q(\theta_k;\eta_k)$ in the product is different. For notational convenience we write $q(\theta_k;\eta_k)$ instead of $q_k(\theta_k;\eta_k)$ when the arguments make it clear which function we are referring to, much as the same symbol p is used to

refer to many different probability distributions without additional indexing. One may additionally assume that the components $q(\theta_k; \eta_k)$ are in a convenient exponential family. We will refer to the use of VB with this choice of the factorization and the exponential family assumption as "MFVB," for "mean field variational Bayes." For example, in the case of MFVB, Ω_η could be a stacked vector of the natural parameters of the exponential families, or the moment parameterization, or perhaps a transformation of these parameters into an unconstrained space (e.g., the entries of log-Cholesky decomposition of a positive definite information matrix). For concrete examples, see Section 4. Although all of our experiments will use MFVB, our results extend to other choices of $\mathcal{Q}$ that satisfy the necessary assumptions.

Recall that we are interested in $\mathbb{E}_{p^x_\alpha}[g(\theta)]$. Our core assumption will be that, for a range of α, the variational distribution provides a good approximation to $\mathbb{E}_{p^x_\alpha}[g(\theta)]$ and its directional derivatives.

Assumption 2.5. *For a given function of interest, $g(\theta)$, a given open set of plausible prior parameters $\mathcal{A}$, and for all $\alpha, \alpha' \in \mathcal{A}$,*

$$\begin{aligned} \mathbb{E}_{q^x_\alpha}[g(\theta)] &\approx \mathbb{E}_{p^x_\alpha}[g(\theta)] \text{ and} \\ \frac{d\mathbb{E}_{q^x_\alpha}[g(\theta)]}{d\alpha^\intercal}(\alpha' - \alpha) &\approx \frac{d\mathbb{E}_{p^x_\alpha}[g(\theta)]}{d\alpha^\intercal}(\alpha' - \alpha). \end{aligned}$$

We will not attempt to be precise about what we mean by the "approximately equal" sign, since we are not aware of any tools for evaluating quantitatively whether Assumption (2.5) holds other than running both VB and MCMC (or some other slow but accurate posterior approximation) and comparing the results. However, VB has been useful in practice to the extent that Assumption (2.5) holds true for at least some parameters of interest. We will evaluate Assumption (2.5) in each of our experiments below by comparing the VB and MCMC posterior approximate means.

Since Assumption (2.5) holds only for a particular choice of $g(\theta)$, it is weaker than the assumption that q^x_α is close to p^x_α in KL divergence, or even that all the posterior means are accurately estimated. For example, as discussed in Appendix B of Giordano et al. [2015] and in Section 10.1.2 of Bishop [2006], a mean field approximation to a multivariate normal posterior produces inaccurate covariances and may have an arbitrarily bad KL divergence from p^x_α, but Assumption (2.5) holds exactly for the location parameters. We discuss the multivariate normal example further in Section (3.1) below.

2.4 Variational Bayes sensitivity

Just as MCMC approximations lend themselves to moment calculations, the variational form of VB approximations lends itself to sensitivity calculations. In this section, as with Theorem (2.1), we derive the sensitivity of VB posterior means to generic perturbations. In Section 3 we will choose particular perturbations to calculate VB prior sensitivity and, through Theorem (2.1), posterior covariances.

Consider a generic class of log-perturbations defined by some function, $f(\theta, t) \in \mathbb{R}$, parameterized by a vector $t \in \mathbb{R}^n$, with $f(\theta, 0) = 0$. Given a choice of $f(\theta, t)$, let us consider the posterior defined by

$$p^x_{\alpha,t}(\theta) = \frac{p(\theta|x)\, p(\theta|\alpha) \exp(f(\theta, t))}{\int p(\theta'|x)\, p(\theta'|\alpha) \exp(f(\theta', t))\, d\theta'}. \tag{8}$$

In the notation of Theorem (2.1), we are taking $\rho(\theta, t) = \log p(\theta|x) + \log p(\theta|\alpha) + f(\theta, t)$. Assuming regularity conditions, given in Appendix D, we can define a variational approximation to this perturbed model:

$$q^x_{\alpha,t}(\theta) := \text{argmin}_{q \in \mathcal{Q}} \left\{ KL\left(q(\theta; \eta) \,||\, p^x_{\alpha,t}(\theta)\right) \right\}. \tag{9}$$

This variational approximation will be a function of t through the optimal parameters $\eta^*(t)$, i.e., $q^x_{\alpha,t}(\theta) = q^x_\alpha(\theta; \eta^*(t))$. For notational convenience, we will define the following quantities.

Definition 2.6. Define the following derivatives of variational expectations evaluated at the optimal parameters:

$$\mathbf{H}_{\eta\eta} := \left.\frac{\partial^2 KL(q(\theta;\eta)||p^x_\alpha(\theta))}{\partial\eta\partial\eta^\intercal}\right|_{\eta=\eta^*} \quad \mathbf{f}_{t\eta} := \left.\frac{\partial \mathbb{E}_{q(\theta;\eta)}[f(\theta,t)]}{\partial t\partial\eta^\intercal}\right|_{\eta=\eta^*,t=0} \quad \mathbf{g}_\eta := \left.\frac{\partial \mathbb{E}_{q(\theta;\eta)}[g(\theta)]}{\partial\eta^\intercal}\right|_{\eta=\eta^*}.$$

Since $g(\theta)$, t, and η are all vectors, the quantities $\mathbf{H}_{\eta\eta}$, $\mathbf{f}_{t\eta}$, and $\mathbf{g}_\eta$ are matrices. With these definitions in hand we can now state:

Theorem 2.7. *Consider a variational approximation $q^x_{\alpha,t}(\theta)$ given in Eq. (9) to the perturbed posterior $p^x_{\alpha,t}(\theta)$ given in Eq. (8). Take the expectation of a posterior expectation of $g(\theta)$ with respect to $q^x_{\alpha,t}(\theta)$. Then, under Assumption (D.1), Assumption (D.2), Assumption (D.3), and Assumption (D.4), and the definitions given in Definition (2.6), we have*

$$\left.\frac{d\mathbb{E}_{q^x_{\alpha,t}}[g(\theta)]}{dt^\intercal}\right|_{t=0} = \mathbf{g}_\eta \mathbf{H}^{-1}_{\eta\eta}\mathbf{f}^\intercal_{t\eta}. \tag{10}$$

A proof and the necessary technical conditions are given in Appendix D. By choosing the appropriate $f(\theta, t)$ and evaluating $\mathbf{f}_{t\eta}$, we can use Theorem (2.7) to calculate the exact sensitivity of VB solutions to nearly any arbitrary local perturbations that satisfy the regularity conditions.

Eq. (10) is formally similar to frequentist sensitivity estimates. For example, the pioneering paper of Cook [1986] contains a formula for assessing the curvature of a marginal likelihood surface [Cook, 1986, Equation 15] that, like our Theorem (2.7), represents the sensitivity as a linear system involving the Hessian of an objective function at its optimum. The geometric interpretation of local robustness suggested by Cook [1986] has been extended to Bayesian settings (see, for example, Zhu et al. [2007, 2011]). In addition to generality, one attractive aspect of their geometric approach is its invariance to parameterization. Investigating geometric interpretations of the present work may be an interesting avenue for future research. Additionally, we note that, much as Neal and Hinton [1998] view VB as a generalization of the expectation-maximization (EM) algorithm, Theorem (2.7) can be understood as a variational generalization of the "structured EM" (SEM) covariance estimate of Meng and Rubin [1991]. We will elaborate on the connection between Theorem (2.7) and SEM in future work.

To close this section, we observe that Theorem (2.7) is the *exact sensitivity* of an *approximate posterior*. That is, when we can solve Theorem (2.7), we have used the choice of the family $\mathcal{Q}$ to simplify the problem enough that its local sensitivity to perturbation has a closed form. In contrast, even when MCMC has converged and is producing draws from the exact posterior, the sample covariance estimate in Eq. (4) represents the *approximate sensitivity* of the *exact posterior*. Theorem (2.7) is then useful to the extent that the VB posterior mean approximates the exact posterior means for all perturbations of interest—that is, to the extent that Assumption (2.5) holds.

3 Calculation and uses of sensitivity

In this section, we briefly discuss practical issues involved in the use of Theorem (2.7). We then apply the results of Section 2 and Assumption (2.5) with particular choices of the perturbation $f(\theta; t)$ to calculate covariances and sensitivity measures for VB estimates. Throughout this section, we will assume that we can apply Theorem (2.1) and Theorem (2.7) unless stated otherwise.

3.1 Covariances for variational Bayes

Consider the mean field approximating family, $\mathcal{Q}_{mf}$, from Section (2.3). It is well known that the resulting marginal variances also tend to be under-estimated even when location parameters are well-estimated (see, e.g., [MacKay, 2003, Wang and Titterington, 2004, Turner and Sahani, 2011, Bishop, 2006, Chapter 10]). Even more obviously, any $q \in \mathcal{Q}_{mf}$ represents as zero the covariance between sub-components of θ that are in different factors of the mean field approximating family. It is therefore unreasonable to expect that $\mathrm{Cov}_{q^x_\alpha}(g(\theta)) \approx \mathrm{Cov}_{p^x_\alpha}(g(\theta))$. However, if Assumption (2.5) holds, we may expect the sensitivity of VB means to certain perturbations to be accurate, and by Theorem (2.1), we expect the corresponding covariances to be accurately estimated by the VB sensitivity. In particular, by taking $f(\theta, t) = t^\intercal g(\theta)$ and under Assumption (2.5), we have

$$\left.\frac{d\mathbb{E}_{q^x_{\alpha,t}}[g(\theta)]}{dt^\intercal}\right|_{t=0} \approx \left.\frac{d\mathbb{E}_{p^x_{\alpha,t}}[g(\theta)]}{dt^\intercal}\right|_{t=0} = \mathrm{Cov}_{p^x_\alpha}(g(\theta)).$$

Thus, we can use Theorem (2.7) to provide an estimate of $\mathrm{Cov}_{p^x_\alpha}(g(\theta))$ that may be superior to $\mathrm{Cov}_{q^x_\alpha}(g(\theta))$. With this motivation in mind, we make the following definition.

Definition 3.1. The *linear response approximation*, $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta))$, to the exact posterior covariance $\mathrm{Cov}_{p^x_\alpha}(g(\theta))$ is given by

$$\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta)) := \mathbf{g}_\eta \mathbf{H}^{-1}_{\eta\eta} \mathbf{g}^\intercal_\eta \approx \mathrm{Cov}_{p^x_\alpha}(g(\theta)). \tag{11}$$

When η^* is a strict local minimum of $KL(q(\theta;\eta)||p^x_\alpha(\theta))$, then $\mathbf{H}_{\eta\eta}$ will be positive definite and symmetric, and, as desired, the covariance estimate $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta))$ will be as well. Since the optimal value of every component of $\mathbb{E}_{q^x_\alpha}[g(\theta)]$ may be affected by the log perturbation $t^\intercal g(\theta)$ in Eq. (8), $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta))$ can estimate non-zero covariances between elements of $g(\theta)$ even when they have been partitioned into separate factors of the mean field approximation.

Note that $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta))$ and $\mathrm{Cov}_{q^x_\alpha}(g(\theta))$ differ only when there are at least some moments of p^x_α that q^x_α fails to accurately estimate. In particular, if q^x_α provided a good approximation to p^x_α for both the first and second moments of $g(\theta)$, then we would have $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta)) \approx \mathrm{Cov}_{q^x_\alpha}(g(\theta))$ since

$$\begin{aligned}
\mathbb{E}_{q^x_\alpha}[g(\theta)] \approx \mathbb{E}_{p^x_\alpha}[g(\theta)] \text{ and} \\
\mathbb{E}_{q^x_\alpha}[g(\theta)\, g(\theta)^\intercal] \approx \mathbb{E}_{p^x_\alpha}[g(\theta)\, g(\theta)^\intercal] &\Rightarrow \mathrm{Cov}_{q^x_\alpha}(g(\theta)) \approx \mathrm{Cov}_{p^x_\alpha}(g(\theta)) \\
\mathbb{E}_{q^x_\alpha}[g(\theta)] \approx \mathbb{E}_{p^x_\alpha}[g(\theta)] &\Rightarrow \mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta)) \approx \mathrm{Cov}_{p^x_\alpha}(g(\theta)).
\end{aligned}$$

Putting these together, we see that

$$\begin{aligned}
\mathbb{E}_{q^x_\alpha}[g(\theta)] \approx \mathbb{E}_{p^x_\alpha}[g(\theta)] \text{ and} \\
\mathbb{E}_{q^x_\alpha}[g(\theta)\, g(\theta)^\intercal] \approx \mathbb{E}_{p^x_\alpha}[g(\theta)\, g(\theta)^\intercal] &\Rightarrow \mathrm{Cov}_{q^x_\alpha}(g(\theta)) \approx \mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta)).
\end{aligned}$$

However, in general, $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta)) \neq \mathrm{Cov}_{q^x_\alpha}(g(\theta))$. In this sense, any discrepancy between $\mathrm{Cov}^{LR}_{q^x_\alpha}(g(\theta))$ and $\mathrm{Cov}_{q^x_\alpha}(g(\theta))$ indicates an inadequacy of the variational approximation for at least the second moments of $g(\theta)$.

Let us consider a simple concrete illustrative example which will demonstrate how $\mathrm{Cov}_{q_\alpha^x}(g(\theta))$ can be a poor approximation to $\mathrm{Cov}_{p_\alpha^x}(g(\theta))$ and how $\mathrm{Cov}_{q_\alpha^x}^{LR}(g(\theta))$ can improve the approximation for some moments but not others. Suppose that the exact posterior is a bivariate normal,

$$p_\alpha^x(\theta) = \mathcal{N}(\theta; \mu, \mathbf{\Sigma}), \tag{12}$$

where $\theta = (\theta_1, \theta_2)^\intercal$, $\mu = (\mu_1, \mu_2)^\intercal$, $\mathbf{\Sigma}$ is invertible, and $\mathbf{\Lambda} := \mathbf{\Sigma}^{-1}$. One may think of μ and $\mathbf{\Sigma}$ as known functions of α and x via Bayes' theorem, for example, as given by a normal-normal conjugate model. Suppose we use the MFVB approximating family

$$\mathcal{Q}_{mf} = \{q(\theta) : q(\theta) = q(\theta_1)\, q(\theta_2)\}.$$

It is not hard to show (see Appendix E) that the optimal VB approximation to p_α^x in the family $\mathcal{Q}_{mf}$ is given by

$$\begin{aligned} q(\theta_1) &= \mathcal{N}\left(\theta_1; \mu_1, \mathbf{\Lambda}_{11}^{-1}\right) \\ q(\theta_2) &= \mathcal{N}\left(\theta_2; \mu_2, \mathbf{\Lambda}_{22}^{-1}\right). \end{aligned}$$

Note that the posterior mean of θ_1 is exactly estimated by the MFVB procedure:

$$\mathbb{E}_{q_\alpha^x}[\theta_1] = \mu_1 = \mathbb{E}_{p_\alpha^x}[\theta_1].$$

However, if $\mathbf{\Sigma}_{12} \neq 0$, then $\mathbf{\Lambda}_{11}^{-1} < \mathbf{\Sigma}_{11}$, and the variance of θ_1 is underestimated. This means that the expectation of θ_1^2 is *not* correctly estimated by the MFVB procedure:

$$\mathbb{E}_{q_\alpha^x}\left[\theta_1^2\right] = \mu_1^2 + \mathbf{\Lambda}_{11}^{-1} \neq \mu_1^2 + \mathbf{\Sigma}_{11} = \mathbb{E}_{p_\alpha^x}\left[\theta_1^2\right].$$

A similar analysis holds for θ_2. Of course, the covariance is also mis-estimated if $\mathbf{\Sigma}_{12} \neq 0$ since, by construction of the MFVB approximation,

$$\mathrm{Cov}_{q_\alpha^x}(\theta_1, \theta_2) = 0 \neq \mathbf{\Sigma}_{12} = \mathrm{Cov}_{p_\alpha^x}(\theta_1, \theta_2).$$

Now let us take $f(\theta, t) = \theta_1 t_1 + \theta_2 t_2$. For all t in a neighborhood of zero, the perturbed posterior given by Eq. (8) remains multivariate normal, so it remains the case that, as a function of t, $\mathbb{E}_{q_{\alpha,t}^x}[\theta_1] = \mathbb{E}_{p_{\alpha,t}^x}[\theta_1]$ and $\mathbb{E}_{q_{\alpha,t}^x}[\theta_2] = \mathbb{E}_{p_{\alpha,t}^x}[\theta_2]$. Consequently, Assumption (2.5) holds with exact equalities when $g(\theta) = \theta$. However, since the second moments are not accurate (irrespective of t), Assumption (2.5) does not hold with exact equalities when $g(\theta) = \left(\theta_1^2, \theta_2^2\right)^\intercal$, nor when $g(\theta) = \theta_1\theta_2$. (Assumption (2.5) may still hold approximately for second moments when $\mathbf{\Sigma}_{12}$ is small.) The fact that Assumption (2.5) holds with exact equalities for $g(\theta) = \theta$ allows us to use Theorem (2.7) and Theorem (2.1) to calculate $\mathrm{Cov}_{q_\alpha^x}^{LR}(g(\theta)) = \mathrm{Cov}_{p_\alpha^x}(g(\theta))$, even though $\mathbb{E}_{p_\alpha^x}[\theta_1\theta_2]$ and $\mathbb{E}_{p_\alpha^x}\left[\left(\theta_1^2, \theta_2^2\right)^\intercal\right]$ are mis-estimated.

In fact, when Assumption (2.5) holds with exact equalities for some θ_i (as in the multivariate Gaussian case just discussed), then the estimated covariance in Eq. (11) for all terms involving θ_i will be exact as well. This is the case for the bivariate normal model above, and described in detail for the general multivariate normal case in Appendix E.

Eq. (11) is a generalization to arbitrary variational approximations of results from Giordano et al. [2015]. Giordano et al. [2015] also demonstrate the effectiveness of Eq. (11) on a range of practical problems. Below, in Section 4, in addition to robustness measures, we will also report the accuracy of Eq. (11) for estimating posterior covariances. We find that, for most parameters of interest, particularly location parameters, $\mathrm{Cov}_{q_\alpha^x}^{LR}(g(\theta))$ provides a good approximation to $\mathrm{Cov}_{p_\alpha^x}(g(\theta))$.

The application of sensitivity measures to VB problems for the purpose of improving covariance estimates has a long history under the name "linear response methods." These methods originated in the statistical physics literature [e.g. Tanaka, 2000, Opper and Saad, 2001] and have been applied to various statistical and machine learning problems [Kappen and Rodriguez, 1998, Tanaka, 1998, Welling and Teh, 2004, Opper and Winther, 2004]. Our work, which builds on our earlier work in Giordano et al. [2015], represents a simplification and generalization of classical linear response methods, and serves to elucidate the relationship between these methods and the local robustness literature. In particular, while Giordano et al. [2015] focused on moment-parameterized exponential families, we derive linear-response covariances for generic variational approximations and connect the linear-response methodology to the Bayesian robustness literature.

A very reasonable approach to the inadequacy of VB covariances is to simply increase the expressiveness of the model class $\mathcal{Q}$ (although, as noted by Turner and Sahani [2011], increased expressiveness does not necessarily lead to better solutions). This is the approach taken by much of the recent VB literature [e.g. Ranganath et al., 2014, Tran et al., 2015a,b, Ranganath et al., 2016, Liu and Wang, 2016]. Though this remains a lively and promising research direction, the use of a more complex class $\mathcal{Q}$ sometimes sacrifices the speed and simplicity that made VB attractive in the first place, and often without the relatively well-understood convergence guarantees of MCMC. We also stress that the current work is not at necessarily at odds with the approach of increasing expressiveness. Sensitivity methods can be a supplement to any VB approximation for which our estimators (which require solving a linear system involving the Hessian of the KL divergence) are tractable.

3.2 Local prior sensitivity for VB

We now turn to estimating prior sensitivity for VB estimates.

Definition 3.2. The *local sensitivity* of $\mathbb{E}_{q^x_\alpha}\left[g\left(\theta\right)\right]$ to prior parameter α is given by

$$\mathbf{S}^q_\alpha := \left.\frac{d\mathbb{E}_{q^x_\alpha}\left[g\left(\theta\right)\right]}{d\alpha}\right|_\alpha.$$

Under Assumption (2.5), $\mathbf{S}^q_\alpha \approx \mathbf{S}_\alpha$. We now to turn to finding a form of $f\left(\theta, t\right)$ for Definition (2.6) to produce $\mathbf{S}^q_\alpha$. Under Assumption (A.1) and Assumption (A.2) in Appendix A, $\log p\left(\theta|\alpha\right)$ is smooth in α and well-defined in a neighborhood of α. Choose a small t having the same dimension as α. Then, up to a constant that captures log-normalizing constants that do not depend on θ (in an abuse of notation, the values of $Constant$ are different from line to line), we have the log posterior for a slightly different α:

$$\begin{aligned}
\log p^x_{\alpha+t}\left(\theta\right) &:= \log p\left(x|\theta\right) + \log p\left(\theta|\alpha+t\right) + Constant \\
&= \log p\left(x|\theta\right) + \log p\left(\theta|\alpha\right) + \left(\log p\left(\theta|\alpha+t\right) - \log p\left(\theta|\alpha\right)\right) + Constant \\
&= \log p^x_\alpha\left(\theta\right) + \left(\log p\left(\theta|\alpha+t\right) - \log p\left(\theta|\alpha\right)\right) + Constant.
\end{aligned}$$

Here, if we take

$$f\left(\theta, t\right) := \log p\left(\theta|\alpha+t\right) - \log p\left(\theta|\alpha\right),$$

we can see that $\mathbf{f}_{t\eta}$ in Definition (2.6) is

$$\mathbf{f}_{t\eta} = \left.\frac{\partial^2 \mathbb{E}_{q^x_\alpha}\left[\log p\left(\theta|\alpha+t\right)\right]}{\partial t \partial \eta^\intercal}\right|_{\eta=\eta^*, t=0} = \left.\frac{\partial}{\partial \eta^\intercal}\mathbb{E}_{q^x_\alpha}\left[\frac{\partial \log p\left(\theta|\alpha\right)}{\partial \alpha}\right]\right|_{\eta=\eta^*, \alpha}. \tag{13}$$

We can thus calculate the variational prior sensitivity:

$$\mathbf{S}_{\alpha}^{q} = \mathbf{g}_{\eta}\mathbf{H}_{\eta\eta}^{-1}\frac{\partial}{\partial\eta^{\intercal}}\mathbb{E}_{q_{\alpha}^{x}}\left[\frac{\partial\log p\left(\theta|\alpha\right)}{\partial\alpha}\right]\Bigg|_{\eta=\eta^{*},\alpha}. \tag{14}$$

Note that—up to the numerical task of calculating the terms in Eq. (14)—$\mathbf{S}_{\alpha}^{q}$ is the exact sensitivity of the variational mean $\mathbb{E}_{q_{\alpha}^{x}}\left[g\left(\theta\right)\right]$. Again, it is the exact sensitivity of an approximate posterior. It is an approximation to $\mathbf{S}_{\alpha}$ to the extent that Assumption (2.5) holds—that is, to the extent that the VB means are good approximations to the exact means. We now use Eq. (14) to reproduce VB versions of some standard robustness measures found in the existing literature.

3.3 Parametric sensitivity

A simple case is when the prior $p\left(\theta|\alpha\right)$ is believed to be in a given parametric family, and we are simply interested in the effect of varying the parametric family's parameters [Basu et al., 1996]. We will refer to this kind of perturbation as a "parametric sensitivity," in contrast to "functional sensitivity," which we discuss in Section (3.4).

For illustration, we first consider a simple example where $p\left(\theta|\alpha\right)$ is in the exponential family, with natural sufficient statistic θ and log normalizer $A\left(\alpha\right)$, and we take $g\left(\theta\right)=\theta$. In this case,

$$\begin{aligned}
\log p\left(\theta|\alpha\right) &= \alpha^{\intercal}\theta - A\left(\alpha\right)\\
\mathbf{f}_{t\eta} &= \frac{\partial}{\partial\eta^{\intercal}}\mathbb{E}_{q_{\alpha}^{x}}\left[\frac{\partial}{\partial\alpha}\left(\alpha^{\intercal}\theta - A\left(\alpha\right)\right)\right]\Bigg|_{\eta=\eta^{*},\alpha}\\
&= \frac{\partial}{\partial\eta^{\intercal}}\mathbb{E}_{q_{\alpha}^{x}}\left[\theta\right] - \frac{\partial}{\partial\eta^{\intercal}}\frac{\partial A\left(\alpha\right)}{\partial\alpha}\Bigg|_{\eta=\eta^{*},\alpha}\\
&= \mathbf{g}_{\eta}.
\end{aligned}$$

Recognizing that, when $\mathbf{f}_{t\eta}=\mathbf{g}_{\eta}$, Eq. (14) is equivalent to Eq. (11), we see that

$$\mathbf{S}_{\alpha}^{q} = \mathrm{Cov}_{q_{\alpha}^{x}}^{LR}\left(\theta\right).$$

In this case, the sensitivity is simply the linear response covariance estimate of the covariance, $\mathrm{Cov}_{q_{\alpha}^{x}}^{LR}\left(\theta\right)$. Following the same reasoning, the exact posterior sensitivity is given by

$$\mathbf{S}_{\alpha} = \mathrm{Cov}_{p_{\alpha}^{x}}\left(\theta\right).$$

Thus, $\mathbf{S}_{\alpha}^{q}\approx\mathbf{S}_{\alpha}$ to the extent that $\mathrm{Cov}_{q_{\alpha}^{x}}^{LR}\left(\theta\right)\approx\mathrm{Cov}_{p_{\alpha}^{x}}\left(\theta\right)$, which again holds to the extent that Assumption (2.5) holds. Note that if we had used a mean field assumption and had tried to use the direct, uncorrected response covariance $\mathrm{Cov}_{q_{\alpha}^{x}}\left(\theta\right)$ to try to evaluate $\mathbf{S}_{\alpha}^{q}$, we would have erroneously concluded that the prior on one component, θ_{k_1}, would not affect the posterior mean of some other component, θ_{k_2}, for $k_2\neq k_1$.

Sometimes it is easy to evaluate the derivative of the log prior even when it is not easy to normalize it. For example, as we show in Appendix F, the LKJ covariance prior [Lewandowski et al., 2009] has a closed-form log expectation when using an inverse Wishart variational approximation. Let $\mathbf{\Sigma}$ be an unknown $K\times K$ information matrix (i.e., the inverse of a covariance matrix that is part of θ). Let

$$\begin{aligned}
q\left(\mathbf{\Sigma}\right) &:= \text{InverseWishart}\left(\mathbf{\Sigma}|\mathbf{\Psi},\nu\right)\\
p\left(\mathbf{\Sigma}|\alpha\right) &\propto \text{LKJ}\left(\mathbf{\Sigma}|\alpha\right),
\end{aligned}$$

where $\mathbf{\Psi}$ is a positive definite scale matrix, ν is the degrees of freedom, and LKJ (α) is a prior on correlation matrices with concentration parameters α. For illustration, we will show how to calculate the local sensitivity to the LKJ concentration parameter. Since we take the partial derivative with respect to α in Eq. (14), we can omit terms from the prior that do not depend on α (e.g. any priors on the diagonal of $\mathbf{\Sigma}$, which we assume are in the constant of proportionality in the definition of $p\left(\mathbf{\Sigma}|\alpha\right)$ and are independent of α). In this case, the variational parameters are $\eta = (\mathbf{\Psi}, \nu)$, with the understanding that we have stacked only the upper-diagonal elements of $\mathbf{\Psi}$, since $\mathbf{\Psi}$ is constrained to be symmetric and η^* must be interior. As we show in Appendix F,

$$\mathbb{E}_{q_\alpha^x}\left[\log p\left(\mathbf{\Sigma}|\alpha\right)\right] = (\alpha - 1)\left(\log|\mathbf{\Psi}| - \psi_K\left(\frac{\nu}{2}\right) - \sum_{k=1}^{K}\log\left(\frac{1}{2}\mathbf{\Psi}_{kk}\right) - K\psi\left(\frac{\nu - K + 1}{2}\right)\right) + Constant,$$

where $Constant$ contains terms that do not depend on α. Here, ψ_K denotes the multivariate digamma function. Consequently, we can evaluate

$$\mathbf{f}_{t\eta} = \frac{\partial}{\partial\eta^\intercal}\mathbb{E}_{q_\alpha^x}\left[\frac{\partial}{\partial\alpha}\log p\left(\mathbf{\Sigma}|\alpha\right)\right] = \frac{\partial}{\partial\eta^\intercal}\left(\log|\mathbf{\Psi}| - \psi_K\left(\frac{n}{2}\right) - \sum_{k=1}^{K}\log\left(\frac{1}{2}\mathbf{\Psi}_{kk}\right) - K\psi\left(\frac{n - K + 1}{2}\right)\right). \tag{15}$$

This derivative has a closed form, though the bookkeeping required to represent an unconstrained parameterization of the matrix $\mathbf{\Psi}$ within η would be tedious. In practice, we evaluate terms like $\mathbf{f}_{t\eta}$ using automatic differentiation tools from the Python `autograd` library [Maclaurin et al., 2015].

Finally, in cases where we cannot evaluate $\mathbb{E}_{q_\alpha^x}\left[\log p\left(\theta|\alpha\right)\right]$ in closed form as a function of η, we can use numerical techniques as described in Section (3.5). We thus view $\mathbf{S}_\alpha^q$ as the exact sensitivity to an approximate KL divergence.

3.4 Functional sensitivity

The parametric perturbations described in Section (3.3) may be insufficiently expressive to describe the full range of plausible priors. Parametric priors are often chosen merely for analytic tractability (e.g., conjugacy) even when much more general functional forms would be subjectively plausible. In this section, we describe the effect of perturbing the prior additively with arbitrary functions, which we refer to as "functional sensitivity." In the present work, we will assume that the practitioner has a particular functional perturbation in mind. However, starting from the results of this section, a VB version of the influence function and worst-case functional perturbation analysis of Gustafson [1996b] follows naturally, though we leave the detailed development of these ideas for future work.

In order to evaluate the effect of changing the prior's functional form, we consider adding to our original prior, $p_0\left(\theta\right)$, a weighted non-negative and λ-integrable measure $u\left(\theta\right)$. After normalizing, we get

$$p\left(\theta|\alpha\right) \quad = \quad \frac{p_0\left(\theta\right) + \alpha u\left(\theta\right)}{\int\left(p_0\left(\theta'\right) + \alpha u\left(\theta'\right)\right)\lambda\left(d\theta'\right)} = \frac{p_0\left(\theta\right) + \alpha u\left(\theta\right)}{1 + \alpha C_u} \text{ for } \alpha \geq 0, \tag{16}$$

where we have defined $C_u := \int u\left(\theta\right)\lambda\left(d\theta\right)$. We will always consider local sensitivity at $\alpha = 0$; i.e., we consider local additive perturbations to $p_0\left(\theta\right)$ using approximations to the posterior distribution with $p\left(\theta|0\right) = p_0\left(\theta\right)$.

Note that $u\left(\theta\right)$ need not be a probability distribution; i.e., we may have $C_u \neq 1$. When $C_u = 1$, then $p\left(\theta|\alpha\right)$ is a mixture between the two distributions $p_0\left(\theta\right)$ and $u\left(\theta\right)$:

$$p\left(\theta|\alpha\right) = (1 - \epsilon)\, p_0\left(\theta\right) + \epsilon u\left(\theta\right) \text{ for } \epsilon := \frac{\alpha}{1 + \alpha}.$$

In this case, the perturbation described in Eq. (16) is known as "ϵ-contamination" [Gustafson, 2000]. Although we differentiate with respect to α instead of ϵ, since $\frac{d\epsilon}{d\alpha}\big|_{\alpha=0} = 1$, the local sensitivity at $p_0(\theta)$ is the same for both parameterizations. We do not always assume that $C_u = 1$, however, since the worst-case perturbation subject to being within a certain distance of $p_0(\theta)$ may not be a probability distribution [Gustafson, 1996b].

In Eq. (16), α is a single non-negative scalar, not a function. In other words, for the purpose of evaluating $\mathbf{S}_\alpha$ we keep $u(\theta)$ fixed, though it is useful to remember that $p(\theta|\alpha)$ in Eq. (16) and $\mathbf{S}_\alpha$ are both functionals of $u(\theta)$. Furthermore, for notational simplicity, will assume that the quantity of interest, $g(\theta)$, is a scalar-valued function when discussing functional sensitivity. To emphasize these points, we make the following definition.

Definition 3.3. When the prior perturbation is of the form Eq. (16), and $g(\theta)$ is a scalar-valued function, we define

$$S_u := \mathbf{S}_\alpha \quad \text{and} \quad S_u^q := \mathbf{S}_\alpha^q.$$

Viewed as a functional of the perturbation $u(\theta)$, derivatives with respect to α become Gateaux derivatives of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ and $\mathbb{E}_{q_\alpha^x}[g(\theta)]$ in the direction of $u(\theta)$ [Huber, 2011, Section 2.5].

Assuming that $u(\theta)$ is sufficiently well behaved, we can apply Corollary (2.3) to Eq. (16).

Proposition 3.4. *Assume a given posterior $p_\alpha^x(\theta)$, a function of interest $g(\theta)$, and an original prior $p_0(\theta)$. Consider the family of prior perturbations given in Eq. (16). Given that Assumption (A.1) holds for our particular $g(\theta)$, then*

$$S_u := \mathbf{S}_\alpha = \mathrm{Cov}_{p_\alpha^x}\left(g(\theta), \frac{u(\theta)}{p_0(\theta)}\right). \tag{17}$$

Proof. Assumption (A.2) will be satisfied since $u(\theta)$ is integrable, and Assumption (A.1) holds by assumption, so we can apply Theorem (2.1). By direct calculation,

$$\left.\frac{\partial \log p(\theta|\alpha)}{\partial \alpha}\right|_{\alpha=0} = \frac{u(\theta)}{p_0(\theta)} - C_u. \tag{18}$$

Since covariances are not affected by mean shifts, the conclusion follows. □

Proposition (3.4) is equivalent to Gustafson [1996a, Result 8], and derived under similar assumptions. We can also derive a variational version of S_u, which we denote S_u^q.

Proposition 3.5. *Assume a given posterior $p_\alpha^x(\theta)$, a function of interest $g(\theta)$, and an original prior $p_0(\theta)$. Consider the family of prior perturbations given in Eq. (16). Given Assumption (D.1), Assumption (D.2), Assumption (D.3), and Assumption (D.4), then*

$$S_u^q \quad = \quad \mathbf{g}_\eta \mathbf{H}_{\eta\eta}^{-1} \frac{\partial}{\partial \eta} \mathbb{E}_{q_\alpha^x}\left[\frac{u(\theta)}{p_0(\theta)}\right]. \tag{19}$$

Proof. Under the given assumptions we can apply Eq. (13), Eq. (14), directly. We can calculate $\mathbf{f}_{t\eta}$ using Eq. (18), noting that $u(\theta)$ is fixed so C_u does not depend on η. □

The necessary assumptions for the VB sensitivity of Proposition (3.5) are not as easily satisfied as those for the exact sensitivity of Proposition (3.4). For example, note that if $q_\alpha^x(\theta)$ has heavier tails than $p_0(\theta)$, then the expectation in Eq. (19) may be infinite, and, correspondingly, the validity of both Assumption (D.3) and Assumption (D.4) will be doubtful.

3.5 Practical considerations and benefits of computing the sensitivity of variational approximations

We briefly discuss practical issues involved in the computation of Eq. (10), which involves calculating the product $\mathbf{g}_\eta \mathbf{H}_{\eta\eta}^{-1}$ (or, equivalently, $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ since $\mathbf{H}_{\eta\eta}$ is symmetric). Calculating $\mathbf{H}_{\eta\eta}$ and solving this linear system can be the most computationally intensive part of computing Eq. (10).

We first note that it can be difficult and time consuming in practice to manually derive and implement second-order derivatives. Even a small programming error can lead to large errors in Theorem (2.7). To ensure accuracy and save analyst time, we evaluated all the requisite derivatives using the Python `autograd` automatic differentiation library [Maclaurin et al., 2015].

Note that the dimension of $\mathbf{H}_{\eta\eta}$ is as large as that of η, the parameters that specify the variational distribution $q(\theta;\eta)$. Many applications of VB employ many latent variables, the number of which may even scale with the amount of data (e.g., the model we examine in Section 4). However, these applications typically have special structure that render $\mathbf{H}_{\eta\eta}$ sparse, allowing the practitioner to calculate $\mathbf{g}_\eta \mathbf{H}_{\eta\eta}^{-1}$ quickly. Consider, for example, a model with "global" parameters, θ_{glob}, that are shared by all the individual datapoint likelihoods, and "local" parameters, $\theta_{loc,n}$, associated with likelihood of a single datapoint indexed by n. By "global" and "local" we mean the likelihood and assumed variational distribution factorize as

$$p(x, \theta_{glob}, \theta_{loc,1}, ..., \theta_{loc,N}) = p(\theta_{glob}) \prod_{n=1}^{N} p(x|\theta_{loc,n}, \theta_{glob})\, p(\theta_{loc,n}|\theta_{glob}) \tag{20}$$

$$q(\theta;\eta) = q(\theta_{glob};\eta_{glob}) \prod_{n=1}^{N} q(\theta_{loc,n};\eta_n) \text{ for all } q(\theta;\eta) \in \mathcal{Q}.$$

In this case, the second derivatives of the variational objective between the parameters for local variables vanish:

$$\text{for all } n \neq m, \frac{\partial^2 KL\left(q(\theta;\eta)\,||p_\alpha^x(\theta)\right)}{\partial \eta_{loc,n} \partial \eta_{loc,m}^\intercal} = 0.$$

The model in Section 4 has such a global or local structure; see Section (4.2) for more details. Additional discussion, including the use of Schur complements to take advantage of sparsity in the log likelihood, can be found in Giordano et al. [2015].

When even calculating or instantiating $\mathbf{H}_{\eta\eta}$ is prohibitively time-consuming, one can use conjugate gradient algorithms to approximately compute $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ [Wright and Nocedal, 1999, Chapter 5]. The advantage of conjugate gradient algorithms is that they use only the Hessian-vector product $\mathbf{H}_{\eta\eta}\mathbf{g}_\eta^\intercal$, which can be computed efficiently using automatic differentiation without ever forming the full Hessian $\mathbf{H}_{\eta\eta}$ (see, for example, the `hessian_vector_product` method of the Python `autograd` package [Maclaurin et al., 2015]). Note that a separate conjugate gradient problem must be solved for each column of $\mathbf{g}_\eta^\intercal$, so if the parameter of interest $g(\theta)$ is high-dimensional it may be faster to pay the price for computing and inverting the entire matrix $\mathbf{H}_{\eta\eta}$. See 4.2 for more discussion of a specific example.

In Theorem (2.7), we require η^* to be at a true local optimum. Otherwise the estimated sensitivities may not be reliable (e.g., the covariance implied by Eq. (11) may not be positive definite). We find that the classical VB coordinate ascent algorithms (Blei et al. [2016, Section 2.4]) and even quasi-second order methods, such as BFGS [e.g. Regier et al., 2015], may not actually find a local optimum unless run for a long time with very stringent convergence criteria. Also, when employing stochastic optimization methods [Hoffman et al., 2013, Ranganath et al., 2016], the estimated optimum is guaranteed to be only near an optimum, rather

than at an optimum. Consequently, we recommend fitting models using second-order Newton trust region methods. When the Hessian is slow to compute directly, as in Section 4, one can use the conjugate gradient trust region method of Wright and Nocedal [1999, Chapter 7], which takes advantage of fast automatic differentiation Hessian-vector products without forming or inverting the full Hessian.

In the current work, a primary reason for calculating a VB version of local sensitivity is to take advantage of VB's speed and scalability. However, $\mathbf{S}_\alpha^q$ and S_u^q may have additional benefits over their corresponding MCMC estimates. First, sample covariance estimates will naturally be subject to Monte Carlo error inherent to MCMC, whereas VB estimates of $\mathbf{S}_\alpha$ are typically either analytically tractable (as in, for example, the LKJ prior of Eq. (15)) or involve only low-dimensional Monte Carlo estimates over known distributions. Examples where only relatively easy Monte Carlo estimates are required for VB include calculating $\mathbf{g}_\eta$ and $\mathbf{f}_{t\eta}$ using samples from q_α^x or intractable integrals with respect to $q(\theta;\eta)$ as in Section 4.

Finally, the sample covariance estimates versions of the functional sensitivity in Eq. (17) may have infinite variance. To calculate Eq. (17) using draws from $p_\alpha^x(\theta)$, we will need sample estimates of $\mathbb{E}_{p_\alpha^x}\left[\frac{g(\theta)u(\theta)}{p_0(\theta)}\right]$. For the variance of this quantity to be finite, we require that the expectation of the square is finite. The expectation of the square is given by

$$\begin{aligned}
\mathbb{E}_{p_\alpha^x}\left[\left(\frac{g(\theta)u(\theta)}{p_0(\theta)}\right)^2\right] &= \int \frac{p_\alpha^x(\theta)}{p_0(\theta)^2}\left(g(\theta)u(\theta)\right)^2\lambda(d\theta) \\
&\propto \int \frac{p(x|\theta)p_0(\theta)}{p_0(\theta)^2}\left(g(\theta)u(\theta)\right)^2\lambda(d\theta) \\
&= \int \frac{p(x|\theta)}{p_0(\theta)}\left(g(\theta)u(\theta)\right)^2\lambda(d\theta).
\end{aligned}$$

This could be infinite if $p_0(\theta)$ has lighter tails than $p(x|\theta)\left(g(\theta)u(\theta)\right)^2$. For example, the variance will be infinite if $p_0(\theta)$ is a normal distribution, $g(\theta)=\theta$, and both $u(\theta)$ and $p(x|\theta)$ have tail behavior in θ like a Student-t distribution. Although Eq. (19) may appear to have a similar problem, since we require

$$\mathbb{E}_{q_\alpha^x}\left[\left(\frac{u(\theta)}{p_0(\theta)}\right)^2\right] = \int \frac{q_\alpha^x(\theta)u(\theta)^2}{p_0(\theta)^2}\lambda(d\theta)$$

to be finite, in Eq. (19) we are able to use importance sampling to reduce the variance. Importance sampling is possible for VB because we have a closed parametric form for the VB posterior, whereas in MCMC we are constrained to use the samples from $p_\alpha^x(\theta)$.

4 Experiments

We demonstrate the techniques above on real data using a logistic regression with random effects, which is an example of a generalized linear mixed model (GLMM) [Agresti and Kateri, 2011, chapter 13]. This data and model have several advantages as an illustration of our methods: the data set is large, the model contains a large number of imprecisely-estimated latent variables (the unknown random effects), the model exhibits the sparsity of $\mathbf{H}_{\eta\eta}$ that is typical in many VB applications, and the results admit an interesting comparison with maximum a posteriori estimates.

All the code necessary to clean the data, run the estimation procedures, and produce the results below can be found in the paper's git repository[1].

[1] `https://github.com/rgiordan/CovariancesRobustnessVBPaper`

4.1 Data and model

We investigated a custom subsample of the 2014 Criteo Labs conversion logs dataset [Criteo Labs, 2014], which contains an obfuscated sample of advertising data collected by Criteo over a period of two months. Each row of the dataset corresponds to an single user click on an online advertisement. For each click, the dataset records a binary outcome variable representing whether or not the user subsequently "converted" (i.e., performed a desired task, such as purchase a product or sign up for a mailing list). Each row contains two timestamps (which we ignore), eight numerical covariates, and nine factor-valued covariates. Of the eight numerical covariates, three contain 30% or more missing data, so we discarded them. We then applied a per-covariate normalizing transform to the distinct values of those remaining. Among the factor-valued covariates, we retained only the one with the largest number of unique values and discarded the others. These data-cleaning decisions were made for convenience. The goal of the present paper is to demonstrate our inference methods, not to draw conclusions about online advertising.

Although the meaning of the covariates has been obfuscated, for the purpose of discussion we will imagine that the single retained factor-valued covariate represents the identity of the advertiser, and the numeric covariates represent salient features of the user and/or the advertiser (e.g., how often the user has clicked or converted in the past, a machine learning rating for the advertisement quality, etc.). As such, it makes sense to model the probability of each row's binary outcome (whether or not the user converted) as a function of the five numeric covariates and the advertiser identity using a logistic GLMM. Specifically, we observe binary conversion outcomes, y_{it} for click i on advertiser t, with probabilities given by observed numerical explanatory variables, x_{it}, each of which are vectors of length $K_x = 5$. Additionally, the outcomes within a given value of t are correlated through an unobserved random effect, u_t, which represents the "quality" of advertiser t, where the value of t for each observation is given by the factor-valued covariate. The random effects u_t are assumed to follow a normal distribution with unknown mean and variance. Formally,

$$\begin{aligned} y_{it}|p_{it}, \beta, u_t, x_{it} &\sim \text{Bernoulli}\left(p_{it}\right), \text{ for } t = 1, ..., T \text{ and } i = 1, ..., N \\ p_{it} &:= \frac{e^{\rho_{it}}}{1 + e^{\rho_{it}}} \quad \text{where} \quad \rho_{it} := x_{it}^T \beta + u_t \\ u_t|\mu, \tau &\sim \mathcal{N}\left(\mu, \tau^{-1}\right). \end{aligned}$$

Consequently, the unknown parameters are $\theta = (\beta^\intercal, \mu, \tau, u_1, ..., u_T)^\intercal$. We use the following priors:

$$\begin{aligned} \mu|\mu_0, \tau_\mu &\sim \mathcal{N}\left(\mu_0, \tau_\mu^{-1}\right) \\ \tau|\alpha_\tau, \beta_\tau &\sim \text{Gamma}\left(\alpha_\tau, \beta_\tau\right) \\ \beta|\beta_0, \tau_\beta, \gamma_\beta &\sim \mathcal{N}\left(\begin{pmatrix} \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix}, \begin{pmatrix} \tau_\beta & \gamma_\beta & \gamma_\beta \\ \gamma_\beta & \ddots & \gamma_\beta \\ \gamma_\beta & \gamma_\beta & \tau_\beta \end{pmatrix}^{-1}\right). \end{aligned}$$

Note that we initially take $\gamma_\beta = 0$ so that the prior information matrix on β is diagonal, though by retaining γ_β as a hyperparameter we will be able to assess the sensitivity to the assumption of a diagonal prior in Section (4.4). The remaining prior values are given in Appendix G. It is reasonable to expect that a modeler would be interested both in the effect of the numerical covariates and in the quality of individual advertisers themselves, so we take the parameter of interest to be $g\left(\theta\right) = (\beta^\intercal, u_1, ..., u_T)^\intercal$.

To produce a dataset small enough to be amenable to MCMC but large and sparse enough to demonstrate our methods, we subsampled the data still further. We randomly chose 5000 distinct advertisers to analyze,

and then subsampled each selected advertiser to contain no more than 20 rows each. The resulting dataset had $N = 61895$ total rows. If we had more observations per advertiser, the "random effects" u_t would have been estimated quite precisely, and the nonlinear nature of the problem would not have been important, obscuring the benefits of using VB versus, say, a maximum a posteriori (MAP) estimate. (We compare our results to a MAP estimator as well as a marginal maximum likelihood estimator in Section (4.3) below.) In typical internet datasets a large amount of data comes from advertisers with few observations each, so our subsample is representative of practically interesting problems.

4.2 Inference and timing

Method	Seconds
MAP (optimum only)	12
VB (optimum only)	57
VB (including sensitivity for β)	104
VB (including sensitivity for β and u)	553
MCMC (Stan)	21066

Table 1: Timing results

We estimated $\mathbb{E}_{p^x_\alpha}[g(\theta)]$ using three techniques: MCMC, MFVB (including the sensitivity tools of Section 3), and the maximum a posteriori (MAP) estimate [Gelman et al., 2014, Chapter 13]. For MCMC, we used Stan [Stan Team, 2015], and for both MFVB and MAP we used our own Python code using `numpy`, `scipy`, and `autograd` [Jones et al., 2001, Maclaurin et al., 2015]. As described in Section (4.3), the MAP estimator did not estimate $\mathbb{E}_{p^x_\alpha}[g(\theta)]$ very well, so we did not attempt to calculate standard deviations or sensitivity measures for the MAP estimator. The summary of the computation time for all these methods is shown in Table (1), with details below.

For the MCMC estimates, we used Stan to draw 5000 MCMC draws (not including warm-up), which took 351 minutes. We estimated all the sensitivities in Section (4.4) using the Monte Carlo version of the covariance in Eq. (3).

For the variational approximation, we use the following mean field exponential family approximations:

$$\begin{aligned}
q(\beta_k) &= \mathcal{N}(\beta_k; \eta_{\beta_k}), \text{ for } k = 1, ..., K_x \\
q(u_t) &= \mathcal{N}(u_t; \eta_{u_t}), \text{ for } t = 1, ..., T \\
q(\tau) &= \text{Gamma}(\tau; \eta_\tau) \\
q(\mu) &= \mathcal{N}(\mu; \eta_\mu) \\
q(\theta) &= q(\tau)\, q(\mu) \prod_{k=1}^{K_x} q(\beta_k) \prod_{t=1}^{T} q(u_t).
\end{aligned}$$

With these choices, evaluating the variational objective requires the following intractable univariate variational expectation:

$$\mathbb{E}_q[\log(1 - p_{it})] \quad = \quad \mathbb{E}_q\left[\log\left(1 - \frac{e^{\rho_{it}}}{1 + e^{\rho_{it}}}\right)\right].$$

We used the re-parameterization trick and four points of Gauss-Hermite quadrature to estimate this integral for each observation. See Appendix G for more details.

We optimized the variational objective using the conjugate gradient Newton's trust region method, `trust-ncg`, of `scipy.optimize`. One advantage of `trust-ncg` is that it performs second-order optimization but requires only Hessian-vector products, which can be computed quickly by `autograd` without constructing the full Hessian. The MFVB fit took 57 seconds, roughly 370 times faster than MCMC with Stan.

With variational parameters for each random effect u_t, $\mathbf{H}_{\eta\eta}$ is a 10014×10014 dimensional matrix. Consequently, evaluating $\mathbf{H}_{\eta\eta}$ directly as a dense matrix using `autograd` would have been prohibitively time-consuming. Fortunately, our model can be decomposed into global and local parameters, and the Hessian term $\mathbf{H}_{\eta\eta}$ in Theorem (2.7) is extremely sparse. In the notation of Section (3.5), take $\theta_{glob} = (\beta^\intercal, \mu, \tau)^\intercal$, take $\theta_{loc,t} = u_t$, and stack the variational parameters as $\eta = \left(\eta_{glob}^\intercal, \eta_{loc,1}, ..., \eta_{loc,T}\right)^\intercal$. The cross terms in $\mathbf{H}_{\eta\eta}$ between the local variables vanish:

$$\frac{\partial^2 KL\left(q\left(\theta;\eta\right)||p_\alpha^x\left(\theta\right)\right)}{\partial\eta_{loc,t_1}\partial\eta_{loc,t_2}} = 0 \text{ for all } t_1 \neq t_2.$$

(Notice that the full likelihood in Appendix G has no cross terms between u_{t_1} and u_{t_2} for $t_1 \neq t_2$.) As the dimension T of the data grows, so does the length of η. However, the dimension of η_{glob} remains constant, and $\mathbf{H}_{\eta\eta}$ remains easy to invert. We show an example of the sparsity pattern of the first few rows and columns of $\mathbf{H}_{\eta\eta}$ in Fig. (1) .

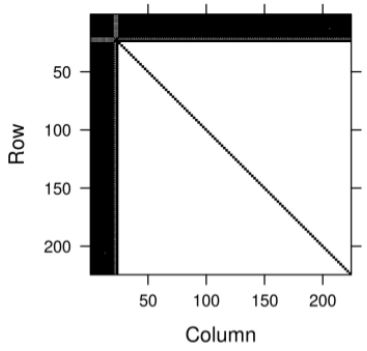

Figure 1: Example sparsity of $\mathbf{H}_{\eta\eta}$ for the logit GLMM model (black indicates non-zero entries)

Taking advantage of this sparsity pattern, we used `autograd` to calculate the Hessian of the KL divergence one group at a time and assembled the results in a sparse matrix using the `scipy.sparse` Python package. Even so, calculating the entire sparse Hessian took 323 seconds, and solving the system $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ using `scipy.sparse.linalg.spsolve` took an additional 173 seconds. This means that the evaluation and inversion of $\mathbf{H}_{\eta\eta}$ was several times more costly than optimizing the variational objective itself. (Of course, the whole procedure remains much faster than running MCMC with Stan.)

We note, however, that instead of the direct approach to calculating $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ one can use the conjugate gradient algorithm of `sp.sparse.linalg.cg` [Wright and Nocedal, 1999, Chapter 5] together with the fast Hessian-vector products of `autograd` to query one column at a time of $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$. On a typical column of $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ in our experiment, calculating the conjugate gradient took only 9.4 seconds (corresponding to 81

Parameter	MCMC	MFVB	MAP	MCMC std. err.	Eff. # of MCMC draws
β_1	1.454	1.447	1.899	0.02067	33
β_2	0.031	0.033	0.198	0.00025	5000
β_3	0.110	0.110	0.103	0.00028	5000
β_4	-0.172	-0.173	-0.173	0.00016	5000
β_5	0.273	0.273	0.280	0.00042	5000
μ	2.041	2.041	3.701	0.04208	28
τ	0.892	0.823	827.724	0.00051	1232
u_{1431}	1.752	1.757	3.700	0.00937	5000
u_{4150}	1.217	1.240	3.699	0.01022	5000
u_{4575}	2.427	2.413	3.702	0.00936	5000
u_{4685}	3.650	3.633	3.706	0.00862	5000

Table 2: Results for the estimation of the posterior means

Hessian-vector products in the conjugate gradient algorithm). Thus, for example, one could calculate the columns of $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ corresponding to the expectations of the global variables β in only $9.4 \times K_x = 46.9$ seconds, which is much less time than it would take to compute the entire $\mathbf{H}_{\eta\eta}^{-1}\mathbf{g}_\eta^\intercal$ for both β and every random effect in u.

For an additional comparison with MFVB, we also calculated the MAP estimate:

$$\theta_{map} := \underset{\theta}{\operatorname{argmax}}\, p_\alpha^x(\theta) = \underset{\theta}{\operatorname{argmax}} \log p_\alpha^x(\theta).$$

We take $g(\theta_{map})$ to be the MAP estimate of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$. The MAP estimator θ_{map}, like the variational approximation $q_\alpha^x(\theta)$, is found by solving an optimization problem that does not require the normalizing constant of $p_\alpha^x(\theta)$. Indeed, the MAP estimator can be seen as equivalent to a MFVB approximation in which every parameter has a degenerate distribution that concentrates at a single point [Neal and Hinton, 1998]. Consequently, the MFVB approximation to posterior means would only be expected to improve on the MAP estimator in cases when there is both substantial uncertainty in some parameters and when this uncertainty, through nonlinear dependence between parameters, affects the values of posterior means. These circumstances obtain in the logistic GLMM model with sparse per-advertiser data, since the random effects u_t will be quite uncertain, and the other posterior means depend on them through the nonlinear logistic function. We calculated the MAP estimator using the same Python code used for the MFVB estimates.

4.3 Posterior approximation results

In this section, we assess the accuracy of the MFVB and MAP estimators as approximations to $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ and $\mathrm{Cov}_{p_\alpha^x}(g(\theta))$, taking the MCMC estimates as ground truth. Although, as discussed in Section (4.1), we are principally interested in the parameters $g(\theta) = (\beta^\intercal, u_1, ..., u_T)^\intercal$, we will report the results for all parameters for completeness. For readability, the tables and graphs show results for a random selection of the components of the random effects u.

4.3.1 Posterior means

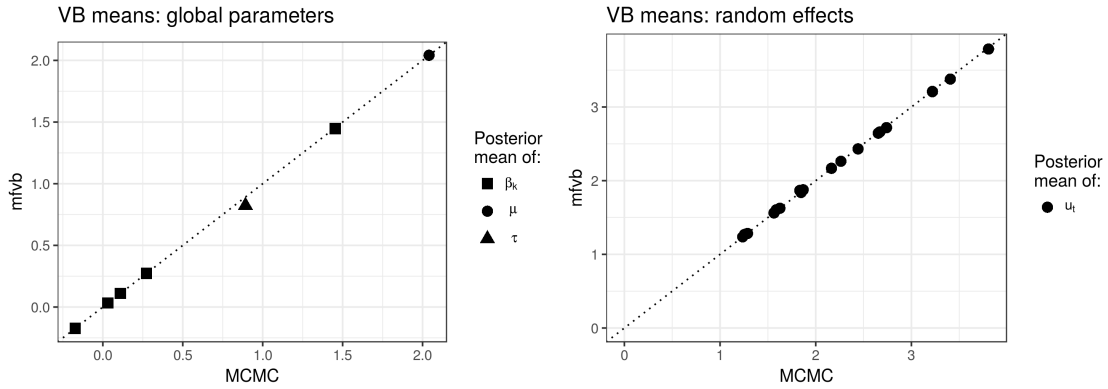

Figure 2: Comparison of MCMC and MFVB means

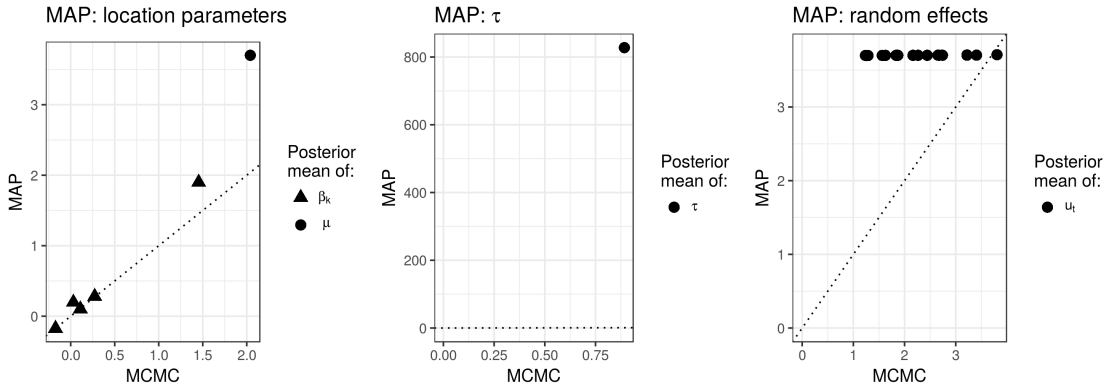

Figure 3: Comparison of MCMC and MAP means

We begin by comparing the posterior means in Table (2), Fig. (2) and Fig. (3). We first note that, despite the long running time for MCMC, the β_1 and μ parameters did not mix well in the MCMC sample, as is reflected in the MCMC standard error and effective number of draws columns of Table (2). The x_{it} data corresponding to β_1 contained fewer distinct values than the other columns of x, which perhaps led to some co-linearity between β_1 and μ in the posterior. This could have caused both poor MCMC mixing and, perhaps, excessive prior sensitivity, as discussed below in Section (4.4). Although we will report the results for both β_1 and μ without further comment, the reader should bear in mind that the MCMC "ground truth" for these two parameters is somewhat suspect.

The results in Table (2) and Fig. (2) show that MFVB does an excellent job of approximating the posterior means in this particular case, even for the random effects u and the related parameters μ and τ. In contrast, the MAP estimator does reasonably well only for certain components of β and does extremely poorly for the random effects parameters. As can be seen in Fig. (3), the MAP estimate dramatically overestimates the information τ of the random effect distribution (that is, it underestimates the variance). As a consequence, it estimates all the random effects to have essentially the same value, leading to mis-estimation of some location parameters, including both μ and some components of β. Because the MAP estimator performed

Parameter	MCMC	LRVB	Uncorrected MFVB
β_1	0.118	0.103	0.005
β_2	0.018	0.018	0.004
β_3	0.020	0.020	0.004
β_4	0.012	0.012	0.004
β_5	0.029	0.030	0.004
μ	0.223	0.192	0.016
τ	0.018	0.033	0.016
u_{1431}	0.663	0.649	0.605
u_{4150}	0.723	0.707	0.662
u_{4575}	0.662	0.649	0.615
u_{4685}	0.610	0.607	0.579

Table 3: Standard deviation results

so poorly at estimating the random effect means, we will not consider it any further.

4.3.2 Posterior covariances

We now assess the accuracy of our estimates of $\mathrm{Cov}_{p_\alpha^x}\left(g\left(\theta\right)\right)$. The results for the diagonals (i.e., the marginal variances) are shown in Table (3) and Fig. (4). We refer to the diagonal of $\mathrm{Cov}_{q_\alpha^x}\left(g\left(\theta\right)\right)$ as the "uncorrected MFVB" estimate, and the diagonal of the linear response covariance estimate $\mathrm{Cov}_{q_\alpha^x}^{LR}\left(g\left(\theta\right)\right)$ of Definition (3.1) as the "LRVB" estimate. To make the scale more meaningful, we report standard deviations rather than variances. The uncorrected MFVB variance estimates $\mathrm{Var}_{q_\alpha^x}\left(\beta\right)$ are particularly inaccurate, but the LRVB variances match the true posterior closely.

In Fig. (5), we compare the off-diagonal elements of $\mathrm{Cov}_{q_\alpha^x}^{LR}\left(g\left(\theta\right)\right)$ and $\mathrm{Cov}_{p_\alpha^x}\left(g\left(\theta\right)\right)$. These covariances are zero, by definition, in the uncorrected MFVB estimates $\mathrm{Cov}_{q_\alpha^x}\left(g\left(\theta\right)\right)$. The left panel of Fig. (5) shows the estimated covariances between the global parameters and all other parameters, including the random effects, and the right panel shows only the covariances amongst the random effects. The LRVB covariances are quite accurate, particularly recalling that the MCMC draws of μ may be inaccurate due to poor mixing.

4.4 Parametric sensitivity results

Finally, we compare the MFVB prior sensitivity measures of Section (3.3) to the covariance-based MCMC sensitivity measures of Section (2.2). Since sensitivity is of practical interest only when it is of comparable order as the posterior uncertainty, we report sensitivities normalized by the appropriate standard deviation. That is, we report $\hat{\mathbf{S}}_\alpha / \sqrt{\mathrm{diag}\left(\hat{\mathrm{Cov}}_{p_\alpha^x}\left(g\left(\theta\right)\right)\right)}$, and $\mathbf{S}_\alpha^q / \sqrt{\mathrm{diag}\left(\mathrm{Cov}_{q_\alpha^x}^{LR}\left(g\left(\theta\right)\right)\right)}$, etc., where $\mathrm{diag}\left(\cdot\right)$ denotes the diagonal vector of a matrix, and the division is element-wise. Note that we use the sensitivity-based variance estimates $\mathrm{Cov}_{q_\alpha^x}^{LR}$, not the uncorrected MFVB estimates $\mathrm{Cov}_{q_\alpha^x}$, to normalize the variational sensitivities. We refer to a sensitivity divided by a standard deviation as a "normalized" sensitivity.

The comparison between the MCMC and MFVB sensitivity measures is shown in Fig. (6). The MFVB and MCMC sensitivities correspond very closely, though the MFVB means appear to be slightly more sensitive to the prior parameters than the MCMC means. This close correspondence should not be surprising. As

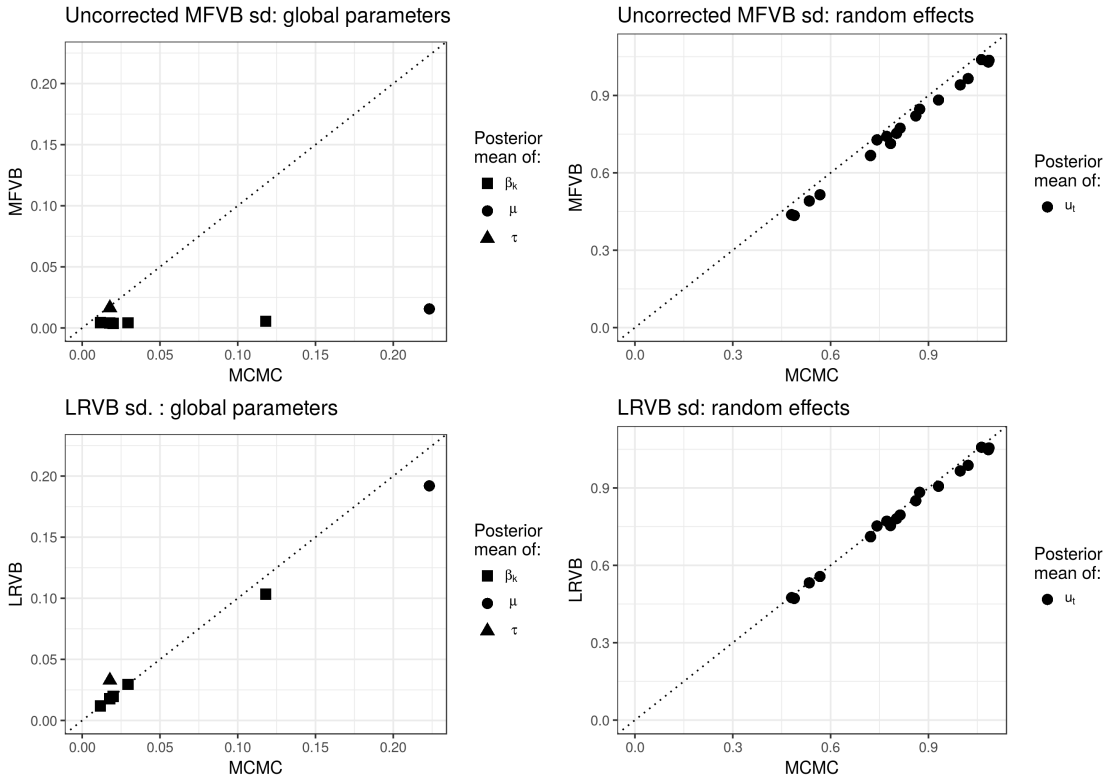

Figure 4: Comparison of MCMC, MFVB, and LRVB standard deviations

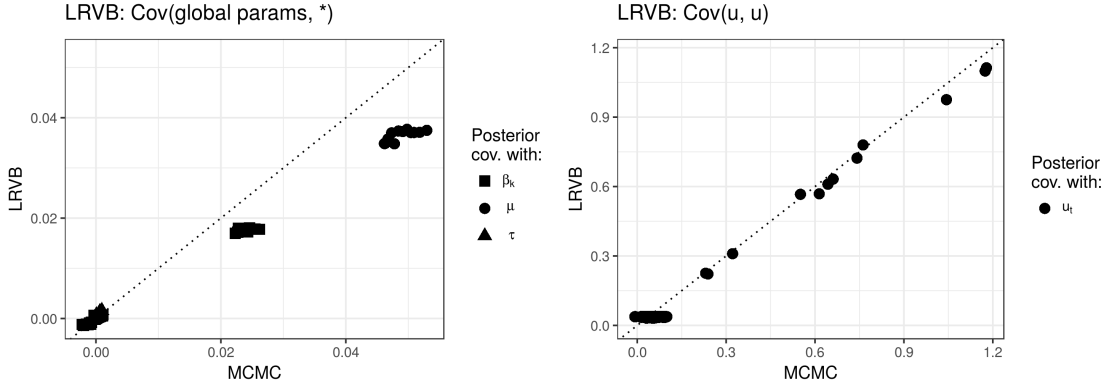

Figure 5: Comparison of MCMC and LRVB off-diagonal covariances

shown in Section (4.3), the MFVB and MCMC posterior means match quite closely. If we assume, reasonably, that they continue to match in a neighborhood of our original prior parameters, then Assumption (2.5) will hold and we would expect $\hat{\mathbf{S}}_\alpha \approx \mathbf{S}_\alpha^q$.

	β_0	τ_β	γ_β	μ_0	τ_μ	α_τ	$\beta\tau$
μ	0.0094	-0.1333	-0.0510	0.0019	-0.3920	0.0058	-0.0048
τ	0.0009	-0.0086	-0.0142	0.0003	-0.0575	0.0398	-0.0328
β_1	0.0089	-0.1464	-0.0095	0.0017	-0.3503	0.0022	-0.0018
β_2	0.0012	-0.0143	-0.0113	0.0003	-0.0516	0.0062	-0.0051
β_3	-0.0035	0.0627	-0.0081	-0.0006	0.1218	-0.0003	0.0002
β_4	0.0018	-0.0037	-0.0540	0.0004	-0.0835	0.0002	-0.0002
β_5	0.0002	0.0308	-0.0695	0.0002	-0.0383	0.0011	-0.0009
u_{1431}	0.0028	-0.0397	-0.0159	0.0006	-0.1169	0.0018	-0.0015
u_{4150}	0.0026	-0.0368	-0.0146	0.0005	-0.1083	0.0022	-0.0018
u_{4575}	0.0028	-0.0406	-0.0138	0.0006	-0.1153	0.0011	-0.0009
u_{4685}	0.0028	-0.0409	-0.0142	0.0006	-0.1163	0.0003	-0.0002

Table 4: MFVB normalized prior sensitivity results

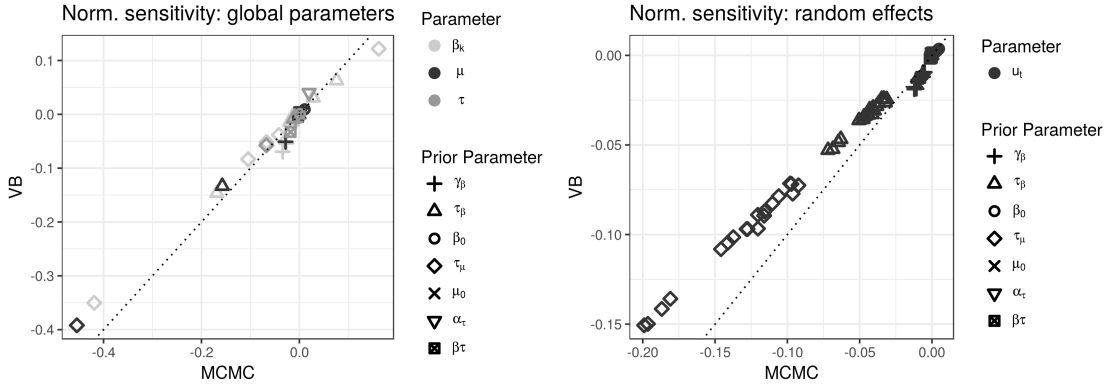

Figure 6: Comparison of MCMC and MFVB normalized parametric sensitivity results

Table (4) shows the detailed MFVB normalized sensitivity results. Each entry is the sensitivity of the VB mean of the row's parameter to the column's prior parameter. One can see that several parameters are quite sensitive to the information parameter prior τ_μ. In particular, $\mathbb{E}_{p_\alpha^x}[\mu]$ and $\mathbb{E}_{p_\alpha^x}[\beta_1]$ are expected to change approximately -0.39 and -0.35 standard deviations, respectively, for every unit change in τ_μ. This size of change could be practically significant (assuming that such a change in τ_μ is subjectively plausible). To investigate this sensitivity further, we re-fit the MFVB model at a range of values of the prior parameter τ_μ, assessing the accuracy of the linear approximation to the sensitivity. The results are shown in Fig. (7). Even for very large changes in τ_μ-—resulting in changes to $\mathbb{E}_{p_\alpha^x}[\mu]$ and $\mathbb{E}_{p_\alpha^x}[\beta_1]$ far in excess of two standard deviations-—the linear approximation holds up reasonably well. Fig. (7) also shows a (randomly selected) random effect to be quite sensitive, though not to a practically important degree relative to its posterior standard deviation. The insensitivity of $\mathbb{E}_{p_\alpha^x}[\beta_2]$ is also confirmed. Of course, the accuracy of the linear approximation cannot be guaranteed to hold as well in general as it does in this particular case, and the quick and reliable evaluation of the linearity assumption without re-fitting the model remains interesting future work.

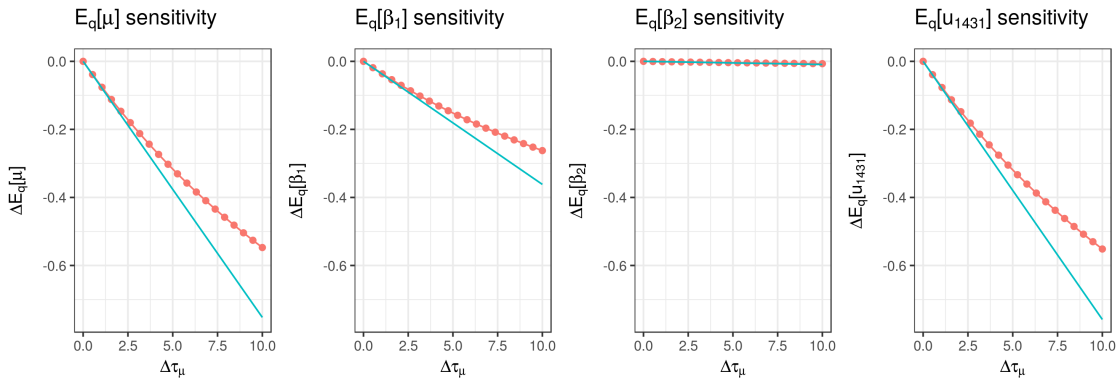

Figure 7: VB sensitivity as measured both by linear approximation and re-fitting

Because we started the MFVB optimization close to the new, perturbed optimum, each new MFVB fit at took only 27.2 seconds on average. Re-estimating the MCMC posterior so many times would have been extremely time-consuming. (Note that importance sampling would be useless for prior parameter changes that moved the posterior so far from the original draws.) The considerable sensitivity of this model to a particular prior parameter, which is perhaps surprising on such a large dataset, illustrates the value of having fast, general tools for discovering and evaluating prior sensitivity. Our framework provides just such a set of tools.

5 Conclusion

By calculating the sensitivity of VB posterior means to model perturbations, we are able to provide two important practical tools for VB posterior approximations: improved variance estimates and measures of prior robustness. When VB models are implemented in software that supports automatic differentiation, our methods are fast, scalable, and require little additional coding beyond the VB objective itself. In our experiments, we were able to calculate accurate posterior means, covariances, and prior sensitivity measures orders of magnitude faster than MCMC.

6 Acknowledgements

Ryan Giordano's research was supported by the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract number DE-AC02- 05CH11231. Tamara Broderick's research was supported in part by a Google Faculty Research Award and the Office of Naval Research under contract/grant number N00014-17-1-2072. This work was also supported in part by a MURI award, W911NF-17-1-0304, from the Army Research Office.

References

A. Agresti and M. Kateri. *Categorical Data Analysis*. Springer, 2011. 4

S. Basu, S. Rao Jammalamadaka, and W. Liu. Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer, 1996. 1, 2.2, 3.3, A, C

J. O. Berger, D. R. Insua, and F. Ruggeri. Robust Bayesian analysis. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2000. 2.2

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. Chapter 10. 2.3, 3.1, E

D. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *arXiv preprint arXiv:1601.00670*, 2016. 1, 2.3, 3.5

R. D. Cook. Assessment of local influence. *Journal of the Royal Statistical Society: Series B*, 28(2):133–169, 1986. 2.4

Criteo Labs. Criteo conversion logs dataset, 2014. URL `http://criteolabs.wpengine.com/downloads/2014-conversion-logs-dataset/`. Downloaded on July 27th, 2017. 4.1

P. Diaconis and D. Freedman. On the consistency of Bayes estimates. *The Annals of Statistics*, 14(1):1–26, 1986. A

B. Efron. Frequentist accuracy of Bayesian estimates. *Journal of the Royal Statistical Society: Series B*, 77 (3):617–646, 2015. 1, 2.2

W. H. Fleming. *Functions of Several Variables*. Addison-Wesley Publishing Company, Inc., 1965. A

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*, volume 2. Chapman & Hall/CRC, 2014. 4.2

R. J. Giordano, T. Broderick, and M. I. Jordan. Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Advances in Neural Information Processing Systems*, pages 1441–1449, 2015. 2.3, 3.1, 3.5, E

P. Gustafson. Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, 91(434):774–781, 1996a. 3.4

P. Gustafson. Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195, 1996b. 1, 3.4, 3.4, A, C

P. Gustafson. Local robustness in Bayesian analysis. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2000. 2.2, 2.2, 3.4

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013. 3.5

P. J. Huber. *Robust Statistics*. Springer, 2011. 2.2, 3.4

D. R. Insua and R. Criado. Topics on the foundations of robust Bayesian analysis. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2000. 2

E. Jones, T. Oliphant, P. Peterson, et al. SciPy: Open source scientific tools for Python, 2001. URL `http://www.scipy.org/`. 4.2, G

H. J. Kappen and F. B. Rodriguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10(5):1137–1156, 1998. 3.1

L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer Science & Business Media, 2012. E

D. Lewandowski, D. Kurowicka, and H. Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009. 3.3, F.1

Q. Liu and D. Wang. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances In Neural Information Processing Systems*, pages 2370–2378, 2016. 3.1

D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Chapter 33. 1, 3.1

D. Maclaurin, D. Duvenaud, and R. P. Adams. Autograd: Effortless gradients in numpy. In *International Conference on Machine Learning 2015 AutoML Workshop*, 2015. 3.3, 3.5, 3.5, 4.2

X. L. Meng and D. B. Rubin. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86(416):899–909, 1991. 2.4

E. Moreno. Global Bayesian robustness for some classes of prior distributions. In D. R. Insua and F. Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2000. 2.2

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer, 1998. 2.4, 4.2

M. Opper and D. Saad. *Advanced Mean Field Methods: Theory and Practice*. MIT press, 2001. 3.1

M. Opper and O. Winther. Variational linear response. In *Advances in Neural Information Processing Systems*, pages 1157–1164, 2004. 3.1

A. B. Owen. *Monte Carlo Theory, Methods and Examples*. 2013. URL `http://statweb.stanford.edu/~owen/mc/`. Accessed November 23rd, 2016. 2.1

C. J. Pérez, J. Martín, and M. J. Rufo. MCMC-based local parametric sensitivity estimations. *Computational Statistics & Data Analysis*, 51(2):823–835, 2006. 2.2, A

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822, 2014. 3.1

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016. 3.1, 3.5

J. Regier, A. Miller, J. McAuliffe, R. Adams, M. Hoffman, D. Lang, D. Schlegel, and M. Prabhat. Celeste: Variational inference for a generative model of astronomical images. In *International Conference on Machine Learning*, pages 2095–2103, 2015. 3.5

M. Roos, T. G. Martins, L. Held, and H. Rue. Sensitivity analysis for Bayesian hierarchical models. *Bayesian Analysis*, 10(2):321–349, 2015. C

Stan Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*, 2015. URL `http://mc-stan.org/`. 4.2

T. Tanaka. Mean-field theory of Boltzmann machine learning. *Physical Review E*, 58(2):2302, 1998. 3.1

T. Tanaka. Information geometry of mean-field approximation. *Neural Computation*, 12(8):1951–1968, 2000. 3.1

D. Tran, D. Blei, and E. M. Airoldi. Copula variational inference. In *Advances in Neural Information Processing Systems*, pages 3564–3572, 2015a. 3.1

D. Tran, R. Ranganath, and D. Blei. The variational Gaussian process. *arXiv preprint arXiv:1511.06499*, 2015b. 3.1

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. In D. Barber, A. T. Cemgil, and S. Chiappa, editors, *Bayesian Time Series Models*. 2011. 1, 3.1, 3.1

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. 1, 2.3

B. Wang and M. Titterington. Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Workshop on Artificial Intelligence and Statistics*, pages 373–380, 2004. 1, 3.1

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *arXiv preprint arXiv:1705.03439*, 2017. 1

M. Welling and Y. W. Teh. Linear response algorithms for approximate inference in graphical models. *Neural Computation*, 16(1):197–221, 2004. 3.1

T. Westling and T. H. McCormick. Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. *arXiv preprint arXiv:1510.08151*, 2015. 1

S. Wright and J. Nocedal. Numerical optimization. *Springer Science*, 35:67–68, 1999. 3.5, 4.2

H. Zhu, J. G. Ibrahim, S. Lee, and H. Zhang. Perturbation selection and influence measures in local influence analysis. *The Annals of Statistics*, 35(6):2565–2588, 2007. 2.4

H. Zhu, J. G. Ibrahim, and N. Tang. Bayesian influence analysis: A geometric approach. *Biometrika*, 98(2): 307–323, 2011. 2.4

Appendices

A Local sensitivity and covariances

In this section we prove Theorem (2.1). It is a straightforward consequence of the Lebesgue dominated convergence theorem. Versions of this theorem have appeared many times before (e.g., Diaconis and Freedman [1986], Basu et al. [1996], Gustafson [1996b], Pérez et al. [2006] to name a few in the robustness literature).

Assumption A.1. *$\rho(\theta, t)$ is continuously differentiable with respect to t, and there exist λ-integrable $f_0(\theta)$ and $f_1(\theta)$ such that $|\exp(\rho(\theta,t))\, g(\theta)| < f_0(\theta)$ and $|\exp(\rho(\theta,t))| < f_1(\theta)$.*

Assumption A.2. *The quantity $\exp(\rho(\theta,t))$ can be normalized with respect to λ, i.e., $0 < \int \exp(\rho(\theta,t))\, \lambda(d\theta) < \infty$.*

When these assumptions hold for all t in a neighborhood of zero, we can prove Theorem (2.1).

Proof. Under Assumption (A.1), we can exchange differentiation and integration in $\int \exp(\rho(\theta,t))\, g(\theta)\, \lambda(d\theta)$ and $\int \exp(\rho(\theta,t))\, \lambda(d\theta)$ by Fleming [1965, Chapter 5-11, Theorem 18], which ultimately depends on the Lebesgue dominated convergence theorem. By Assumption (A.2), $\mathbb{E}_{\rho(\theta,t)}[g(\theta)]$ is well-defined in a neighborhood of $t = 0$, and

$$\frac{\partial \exp(\rho(\theta,t))}{\partial t} = \frac{\partial \rho(\theta,t)}{\partial t}\rho(\theta,t) \quad \lambda\text{-almost everywhere.}$$

Armed with these facts, we can directly compute

$$\begin{aligned}
\left.\frac{\partial \mathbb{E}_{\rho(\theta,t)}[g(\theta)]}{\partial t^\intercal}\right|_{t=0} &= \left.\frac{\partial}{\partial t^\intercal}\,\frac{\int g(\theta)\exp(\rho(\theta,t))\,\lambda(d\theta)}{\int \exp(\rho(\theta,t))\,\lambda(d\theta)}\right|_{t=0} \\
&= \left.\frac{\frac{\partial}{\partial t^\intercal}\int g(\theta)\exp(\rho(\theta,t))\,\lambda(d\theta)}{\int \exp(\rho(\theta,t))\,\lambda(d\theta)}\right|_{t=0} - \mathbb{E}_{\rho(\theta,t)}[g(\theta)]\left.\frac{\frac{\partial}{\partial t^\intercal}\int \exp(\rho(\theta,t))\,\lambda(d\theta)}{\int \exp(\rho(\theta,t))\,\lambda(d\theta)}\right|_{t=0} \\
&= \left.\frac{\int g(\theta)\,\frac{\partial\rho(\theta,t)}{\partial t^\intercal}\exp(\rho(\theta,t))\,\lambda(d\theta)}{\int \exp(\rho(\theta,t))\,\lambda(d\theta)}\right|_{t=0} - \mathbb{E}_{\rho(\theta,0)}[g(\theta)]\left.\frac{\int \frac{\partial\rho(\theta,t)}{\partial t^\intercal}\exp(\rho(\theta,t))\,\lambda(d\theta)}{\int \exp(\rho(\theta,t))\,\lambda(d\theta)}\right|_{t=0} \\
&= \mathrm{Cov}_{\rho(\theta,0)}\left(g(\theta), \left.\frac{\partial\rho(\theta,t)}{\partial t}\right|_{t=0}\right).
\end{aligned}$$

□

B Comparison with MCMC importance sampling

In this section, we show that using importance sampling with MCMC samples to calculate the local sensitivity Eq. (2) is precisely equivalent to using the same MCMC samples to estimate the covariance in Eq. (1) directly. For this section, will suppose that Assumption (A.1) and Assumption (A.2) hold. Further suppose,

without loss of generality, we have samples θ_i drawn iid from the normalized version of $\exp(\rho(\theta,t))$:

$$\begin{aligned} p(\theta|t) &:= \frac{\exp(\rho(\theta,t))}{\int \exp(\rho(\theta',t))\,\lambda(d\theta')} \\ \theta_n &\overset{iid}{\sim} p(\theta|0)\,, \text{ for } n = 1,...,N \\ \mathbb{E}_{p(\theta|0)}[g(\theta)] &\approx \frac{1}{N}\sum_{n=1}^{N} g(\theta_n)\,. \end{aligned}$$

Define the normalizing constant

$$C_t := \int \exp(\rho(\theta',t))\,\lambda(d\theta')\,.$$

If we could calculate C_t, we could use the following importance sampling estimate for $\mathbb{E}_{p(\theta|t)}[g(\theta)]$:

$$\begin{aligned} \mathbb{E}_{p(\theta|t)}[g(\theta)] &\approx \frac{1}{N}\sum_{n=1}^{N} w_n g(\theta_n) \\ w_n &:= \frac{p(\theta_n|t)}{p(\theta_n|0)} \\ &= \exp(\rho(\theta_n,t) - \rho(\theta_n,0) + \log C_0 - \log C_t)\,. \end{aligned}$$

Differentiating the weights,

$$\begin{aligned} \frac{\partial w_n}{\partial t} &= w_n\left(\frac{\partial \rho(\theta_n,t)}{\partial t} - \frac{\partial \log C_t}{\partial t}\right) \\ &= w_n\left(\frac{\partial \rho(\theta_n,t)}{\partial t} - \frac{1}{C_t}\int \exp(\rho(\theta,t))\,\frac{\partial \rho(\theta,t)}{\partial t}d\theta\right) \\ &= w_n\left(\frac{\partial \rho(\theta_n,t)}{\partial t} - \mathbb{E}_{p(\theta|t)}\left[\frac{\partial \rho(\theta,t)}{\partial t}\right]\right). \end{aligned}$$

It follows that

$$\frac{\partial}{\partial t}\frac{1}{N}\sum_{n=1}^{N} w_n g(\theta_n)\bigg|_{t=0} = \frac{1}{N}\sum_{n=1}^{N}\left(\frac{\partial \rho(\theta_n,t)}{\partial t}\bigg|_{t=0} - \mathbb{E}_{p(\theta|0)}\left[\frac{\partial \rho(\theta,t)}{\partial t}\bigg|_{t=0}\right]\right) g(\theta_n)\,,$$

which is precisely the MCMC sample estimate of the covariance given by Theorem (2.1).

C Our use of the terms "sensitivity" and "robustness"

In this section we clarify our usage of the terms "robustness" and "sensitivity." The quantity $\mathbf{S}_\alpha^\intercal \Delta\alpha$ measures the *sensitivity* of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ to perturbations in the direction $\Delta\alpha$. Intuitively, as sensitivity increases, robustness decreases, and, in this sense, sensitivity and robustness are opposites of one another. However, we emphasize that sensitivity is a clearly defined, measurable quantity and that robustness is a subjective judgment informed by sensitivity, but also by many other less objective considerations.

Suppose we have calculated the sensitivity to changes in the direction $\Delta\alpha$ from Eq. (2) and found that it has a particular value. To determine whether our model is robust, we must additionally decide

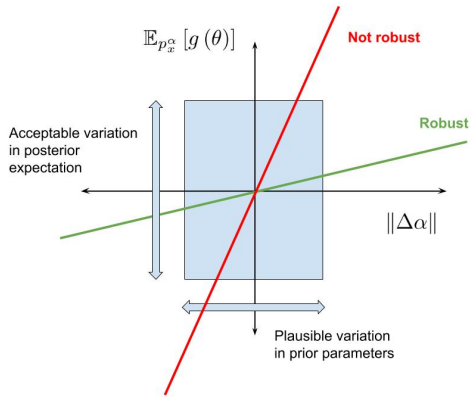

Figure 8: The relationship between robustness and sensitivity

1. How large of a change in the prior, $||\Delta\alpha||$, is plausible, and
2. How large of a change in $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ is important.

The set of plausible prior values necessarily remains a subjective decision[2]. Whether or not a particular change in $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ is important depends on the ultimate use of the posterior mean. For example, the posterior standard deviation can be a guide: if the prior sensitivity is swamped by the posterior uncertainty then it can be neglected when reporting our subjective uncertainty about $g(\theta)$, and the model is robust. Similarly, even if the prior sensitivity is much larger than the posterior standard deviation but small enough that it would not affect any actionable decision made on the basis of the value of $\mathbb{E}_{p_\alpha^x}[g(\theta)]$, then the model is robust. Intermediate values remain a matter of judgment. A illustration of the relationship between sensitivity and robustness is shown in Fig. (8).

Finally, we note that if $\mathcal{A}$ is small enough that $\mathbb{E}_{p_\alpha^x}[g(\theta)]$ is roughly linear in α for $\alpha \in \mathcal{A}$, then calculating Eq. (2) for all $\Delta\alpha \in \mathcal{A} - \alpha$ and finding the worst case can be thought of as a first-order approximation to a global robustness estimate. Often, this linearity assumption is not plausible except for very small $\mathcal{A}$, particularly for function-valued perturbations, as will be seen in Section 4. This is a weakness inherent to the local robustness approach. Nevertheless, even when the perturbations are valid only for a small $\mathcal{A}$, these easily-calculable measures can still provide valuable intuition about the potential modes of failure for a model.

When the prior has many parameters (i.e., α is high dimensional) many authors [e.g. Basu et al., 1996, Gustafson, 1996b, Roos et al., 2015] attempt to summarize the high-dimensional vector $\mathbf{S}_\alpha$ in a single easily reported number such as

$$\mathbf{S}_\alpha^{sup} := \sup_{\alpha:\|\alpha\|\leq 1} \mathbf{S}_\alpha.$$

Although this summary has obvious merits, in this work we do not attempt to summarize $\mathbf{S}_\alpha$ in this way for several reasons. First of all, the unit ball $\|\alpha\| \leq 1$ (as in Basu et al. [1996]) may not make sense as a subjective description of the range of plausible variability of $p(\theta|\alpha)$—why should the off-diagonal term

[2]This decision can be cast in a formal decision theoretic framework based on a partial ordering of subjective beliefs. [Insua and Criado, 2000]

of a Wishart prior plausibly vary as widely as the mean of some other parameter, when the two might not even have the same units? This problem might be easily remedied by choosing an appropriate scaling of the parameters and thereby making the unit ball an appropriate range for the problem at hand, but the right scaling will vary from problem to problem and necessarily be a somewhat subjective choice, so we refrain from taking a stand on this decision.

D Variational Bayes sensitivity derivations and assumptions

Using the notation from Section (2.4), we will make the following assumptions:

Assumption D.1. *The posterior in Eq. (8) is well-defined for t in a neighborhood of zero; i.e., there exists some $\delta > 0$ such that*

$$\int p(\theta|x)\, p(\theta|\alpha) \exp(f(\theta, t))\, d\theta \quad < \quad \infty, \forall t \in \{t : \|t\|_2 < \delta\}.$$

Assumption D.2. *There exists a local minimum, η^*, of $KL(q(\theta;\eta)\,||p^x_\alpha(\theta))$ in Eq. (6), such that η^* is interior to Ω_η.*

Assumption D.3. *The KL divergence, $KL\left(q(\theta;\eta)\,||p^x_{\alpha,t}(\theta)\right)$ is twice differentiable and strictly convex for η in a neighborhood of the optimal η^* and t in a neighborhood of zero.*

Assumption D.4. *The optimum $\eta^*(t)$ of $KL\left(q(\theta;\eta)\,||p^x_{\alpha,t}(\theta)\right)$ is a continuously differentiable function of t in a neighborhood of $\eta^* = \eta^*(0)$.*

We now prove Theorem (2.7).

Proof. Under Assumption (D.1), we can define a variational approximation to the tilted likelihood $p^x_{\alpha,t}(\theta)$ that is a function of t:

$$q^x_{\alpha,t}(\theta) \quad := \quad \text{argmin}_{q\in\mathcal{Q}} \left\{ KL\left(q(\theta;\eta)\,||p^x_{\alpha,t}(\theta)\right)\right\}.$$

For notational convenience, we will define

$$KL(\eta, t) \quad := \quad KL\left(q(\theta;\eta)\,||p^x_{\alpha,t}(\theta)\right).$$

Since by Assumption (D.2) $\eta^*(t)$ is both optimal and interior for all t in a neighborhood of zero, and by Assumption (D.3) $KL(\eta, t)$ is smoothly differentiable in η, the first order conditions of the optimization problem Eq. (6) give:

$$\left.\frac{\partial KL(\eta, t)}{\partial \eta}\right|_{\eta=\eta^*(t)} \quad = \quad 0.$$

By Assumption (D.4) and Assumption (D.3), we can take the total derivative of this vector-valued first order condition with respect to t, getting

$$\left.\frac{\partial^2 KL(\eta, t)}{\partial\eta\partial\eta^T}\right|_{\eta=\eta^*(t)} \frac{d\eta^*(t)}{dt^T} + \left.\frac{\partial^2 KL(\eta, t)}{\partial\eta\partial t^T}\right|_{\eta=\eta^*(t)} \quad = \quad 0.$$

The strict convexity of $KL(\eta, t)$ around η^* in Assumption (D.3) requires that $\frac{\partial^2 KL(\eta,t)}{\partial\eta\partial\eta^T}\Big|_{\eta=\eta^*(t)}$ be invertible, so by evaluating at $t = 0$ and solving we find that

$$\left.\frac{d\eta^*(t)}{dt^T}\right|_{t=0} = -\left(\frac{\partial^2 KL(\eta,t)}{\partial\eta\partial\eta^T}\right)^{-1}\left.\frac{\partial^2 KL(\eta,t)}{\partial\eta\partial t^T}\right|_{\eta=\eta^*,t=0}.$$

Noting that $\mathbb{E}_{q^x_{\alpha,t}}[g(\theta)]$ is a function of $\eta^*(t)$, by Assumption (D.4) we have

$$\left.\frac{d\mathbb{E}_{q^x_{\alpha,t}}[g(\theta)]}{dt^T}\right|_{t=0} = \frac{\partial\mathbb{E}_{q^x_\alpha}[g(\theta)]}{\partial\eta}\left.\frac{d\eta^*(t)}{dt^T}\right|_{\eta=\eta^*,t=0}.$$

Finally, we observe that

$$\begin{aligned} KL(q(\theta;\eta)||p^x_\alpha(\theta;t)) &= \mathbb{E}_{q^x_\alpha}[\log q(\theta;\eta) - \log p(x|\theta) - \log p(\theta|\alpha) - f(\theta,t)] + Constant \Rightarrow \\ \left.\frac{\partial^2 KL(\eta,t)}{\partial\eta\partial t^T}\right|_{t=0} &= -\left.\frac{\partial^2\mathbb{E}_{q^x_\alpha}[f(\theta,t)]}{\partial\eta\partial t^\intercal}\right|_{\eta=\eta^*,t=0}. \end{aligned}$$

Here, the term $Constant$ contains quantities that do not depend on η. Plugging in gives the desired result. □

E Exactness of multivariate normal posterior means

Here, we show that the MFVB estimate of the posterior means of a multivariate normal with known covariance is exact and that, as an immediate consequence, the linear response covariance recovers the exact posterior covariance, i.e., $\mathrm{Cov}^{LR}_{q^x_\alpha}(\theta) = \mathrm{Cov}_{p^x_\alpha}(\theta)$.

Suppose we are using MFVB to approximate a non-degenerate multivariate normal posterior, i.e.,

$$p^x_\alpha(\theta) = \mathcal{N}(\theta;\mu,\mathbf{\Sigma})$$

for full-rank $\mathbf{\Sigma}$. This posterior arises, for instance, given a multivariate normal likelihood $p(x|\mu) = \prod_{n=1:N}\mathcal{N}(x_n|\theta,\mathbf{\Sigma}_x)$ with known covariance $\mathbf{\Sigma}_x$ and a conjugate multivariate normal prior on the unknown mean parameter θ. Additionally, even when the likelihood is non-normal or the prior is not conjugate, the posterior may be closely approximated by a multivariate normal distribution when a Bayesian central limit theorem can be applied [Le Cam and Yang, 2012, Chapter 8]. We will consider an MFVB approximation to $p^x_\alpha(\theta)$. Specifically, let the elements of the vector $\theta \in \mathbb{R}^K$ be given by θ_k, for $k = 1, ..., K$, and take the MFVB normal approximation with means m_k and variances v_k:

$$\mathcal{Q} = \left\{ q(\theta) : q(\theta) = \prod_{k=1}^{K}\mathcal{N}(\theta_k; m_k, v_k) \right\}.$$

In the notation of Eq. (7), we have $\eta_k = (m_k, v_k)^\intercal$. The optimal variational parameters are given by $\eta^*_k = (m^*_k, v^*_k)^\intercal$.

Lemma E.1. *Let $p_\alpha^x(\theta) = \mathcal{N}(\theta; \mu, \mathbf{\Sigma})$ for full-rank $\mathbf{\Sigma}$; let $\mathcal{Q} = \left\{ q(\theta) : q(\theta) = \prod_{k=1}^K \mathcal{N}(\theta_k; m_k, v_k) \right\}$ be the mean field approximating family; and let the optimal parameter $\eta^* = (m^*, v^*)$ solve*

$$\eta^* = \underset{\eta: q(\theta;\eta) \in \mathcal{Q}}{\operatorname{argmin}} \, KL\left(q(\theta;\eta) \,||p_\alpha^x(\theta)\right).$$

Then $m^* = \mu$, i.e., the variational posterior means exactly recover the exact posterior means and η^* is interior.

Proof. Let $\operatorname{diag}(v)$ denote the $K \times K$ matrix with v on the diagonal and zero elsewhere. Using the fact that the entropy of a univariate normal distribution with variance v is $\frac{1}{2}\log v$ plus a constant, the variational objective Eq. (6) is given by

$$\begin{aligned}
KL\left(q(\theta;\eta) \,||p_\alpha^x(\theta)\right) &= \mathbb{E}_{q(\theta;\eta)}\left[-\frac{1}{2}(\theta-\mu)^\intercal \mathbf{\Sigma}^{-1}(\theta-\mu)\right] + \frac{1}{2}\sum_k \log v_k + C \\
&= -\frac{1}{2}\text{trace}\left(\mathbf{\Sigma}^{-1}\mathbb{E}_{q(\theta;\eta)}\left[\theta\theta^\intercal\right]\right) + \mu^\intercal\mathbf{\Sigma}^{-1}\mathbb{E}_{q(\theta;\eta)}\left[\theta\right] + \frac{1}{2}\sum_k \log v_k + C \\
&= -\frac{1}{2}\text{trace}\left(\mathbf{\Sigma}^{-1}\left(mm^\intercal + \operatorname{diag}(v)\right)\right) + \mu^\intercal\mathbf{\Sigma}^{-1}m + \frac{1}{2}\sum_k \log v_k + C \\
&= -\frac{1}{2}\text{trace}\left(\mathbf{\Sigma}^{-1}\operatorname{diag}(v)\right) - \frac{1}{2}m^\intercal\mathbf{\Sigma}^{-1}m + \mu^\intercal\mathbf{\Sigma}^{-1}m + \frac{1}{2}\sum_k \log v_k + C.
\end{aligned}$$

The first order condition for the optimal m^* is then

$$\begin{aligned}
\left.\frac{\partial KL\left(q(\theta;\eta) \,||p_\alpha^x(\theta)\right)}{\partial m}\right|_{m=m^*, v=v^*} &= 0 \Rightarrow \\
-\mathbf{\Sigma}^{-1}m^* + \mathbf{\Sigma}^{-1}\mu &= 0 \Rightarrow \\
m^* &= \mu.
\end{aligned}$$

The optimal variances follow similarly:

$$\begin{aligned}
\left.\frac{\partial KL\left(q(\theta;\eta) \,||p_\alpha^x(\theta)\right)}{\partial v_k}\right|_{m=m^*, v=v^*} &= 0 \Rightarrow \\
-\frac{1}{2}\left(\mathbf{\Sigma}^{-1}\right)_{kk} + \frac{1}{2}\frac{1}{v_k^*} &= 0 \Rightarrow \\
v_k^* &= \frac{1}{\left(\mathbf{\Sigma}^{-1}\right)_{kk}}.
\end{aligned}$$

Clearly, both m^* and v^* are interior. The same result can be derived via the variational coordinate ascent updates (Bishop [2006, Section 10.1.2] and Giordano et al. [2015, Appendix B]). □

Next, we show that Lemma E.1 holds for all perturbations in Eq. (8) of the form $f(\theta, t) = t^\intercal\theta$ and that the assumptions necessary for application of Theorem (2.7) are satisfied.

Lemma E.2. *Under the conditions of Lemma E.1, let $p_{\alpha,t}^x(\theta)$ be defined from Eq. (8) with $f(\theta, t) = t^\intercal\theta$ and let $q_{\alpha,t}^x(\theta)$ satisfy Eq. (9) with the approximating family given in Lemma E.1. Then $\mathbb{E}_{q_{\alpha,t}^x}[\theta] = \mathbb{E}_{p_{\alpha,t}^x}[\theta]$ and Assumption (D.1), Assumption (D.2), Assumption (D.3), and Assumption (D.4) are satisfied for all t in a neighborhood of zero.*

Proof. By standard properties of the multivariate normal distribution, $p^x_{\alpha,t}(\theta)$ is multivariate normal for any t in a neighborhood of zero when $f(\theta, t) = t^\intercal\theta$. This property holds since θ is a sufficient statistic of the multivariate normal distribution, and the corresponding natural parameter of $p^x_\alpha(\theta)$, $\mathbf{\Sigma}^{-1}\mu$, is interior when $\mathbf{\Sigma}$ is full-rank. Assumption (D.1) follows because the non-degenerate multivariate normal distribution is normalizable, and, by Lemma E.1, we have that $\mathbb{E}_{q^x_{\alpha,t}}[\theta] = \mathbb{E}_{p^x_{\alpha,t}}[\theta]$ and Assumption (D.2) holds. Assumption (D.3) can be confirmed by directly inspecting the form of the $\mathbb{E}_{q^x_\alpha}\left[\log p^x_{\alpha,t}(\theta)\right]$ term in the perturbed KL divergence:

$$\begin{aligned}\mathbb{E}_{q^x_\alpha}\left[\log p^x_{\alpha,t}(\theta)\right] = &-\frac{1}{2}\text{trace}\left(\mathbf{\Sigma}^{-1}\text{diag}(v)\right) - \frac{1}{2}m^\intercal\mathbf{\Sigma}^{-1}m + \mu^\intercal\mathbf{\Sigma}^{-1}m \\ &- \frac{1}{2}\sum_k \log v_k + t^\intercal m + Constant.\end{aligned}$$

Here, $Constant$ contains quantities that do not depend on the variational distribution. Finally, Assumption (D.4) holds because, as shown in Lemma E.1, η^* is a smooth function of the natural exponential family parameters of $p^x_{\alpha,t}(\theta)$, which are in turn smooth functions of t. □

It now follows that the linear response variational covariance exactly reproduces the exact posterior covariance.

Theorem E.3. *Under the conditions of Lemma E.1,* $\text{Cov}^{LR}_{q^x_\alpha}(\theta) = \text{Cov}_{p^x_\alpha}(\theta)$.

Proof. From Lemma E.2 we have that, for all t in a neighborhood of zero, $\mathbb{E}_{q^x_{\alpha,t}}[\theta] = \mathbb{E}_{p^x_{\alpha,t}}[\theta]$. It follows that

$$\left.\frac{d\mathbb{E}_{q^x_{\alpha,t}}[\theta]}{dt}\right|_{t=0} = \left.\frac{d\mathbb{E}_{p^x_{\alpha,t}}[\theta]}{dt}\right|_{t=0}.$$

Since, as argued in Lemma E.2, $p^x_{\alpha,t}(\theta)$ is multivariate normal for t in a neighborhood of zero, Assumption (A.1) and Assumption (A.2) are satisfied, so we can apply Theorem (2.1) to get

$$\left.\frac{d\mathbb{E}_{p^x_{\alpha,t}}[\theta]}{dt}\right|_{t=0} = \text{Cov}_{p^x_\alpha}(\theta).$$

Finally, since by Lemma E.2 the conditions for application of Theorem (2.7) are satisfied with $g(\theta) = \theta$, we have

$$\left.\frac{d\mathbb{E}_{q^x_{\alpha,t}}[\theta]}{dt}\right|_{t=0} = \text{Cov}^{LR}_{q^x_\alpha}(\theta).$$

It follows directly that $\text{Cov}^{LR}_{q^x_\alpha}(\theta) = \text{Cov}_{p^x_\alpha}(\theta)$. □

F LKJ Priors for Covariance Matrices in Mean Field Variational Inference

In this section we briefly derive closed-form expressions for using an LKJ prior with a Wishart variational approximation.

Proposition F.1. *Let $\mathbf{\Sigma}$ be a $K \times K$ positive definite covariance matrix. Define the $K \times K$ matrix $\mathbf{S}$ such that*

$$\mathbf{S}_{ij} = \begin{cases} \sqrt{\mathbf{\Sigma}_{ij}} & \text{if } i = j \\ 0 & \text{otherwise.} \end{cases}$$

Define the correlation matrix R as

$$\mathbf{R} = \mathbf{S}^{-1}\mathbf{\Sigma}\mathbf{S}^{-1}.$$

Define the LKJ prior on $\mathbf{R}$ with concentration parameter ξ [Lewandowski et al., 2009]:

$$p\left(\mathbf{R}|\xi\right) \propto |\mathbf{R}|^{\xi-1}.$$

Let $q\left(\mathbf{\Sigma}|\mathbf{V}^{-1}, \nu\right)$ be an inverse Wishart distribution with matrix parameter $\mathbf{V}^{-1}$ and degrees of freedom ν. Then

$$\mathbb{E}_q\left[\log|\mathbf{R}|\right] = \log\left|\mathbf{V}^{-1}\right| - \psi_K\left(\frac{\nu}{2}\right) - \sum_{k=1}^{K}\log\left(\left(\mathbf{V}^{-1}\right)_{kk}\right) - K\psi\left(\frac{\nu-K+1}{2}\right) + Constant$$

$$\mathbb{E}_q\left[\log p\left(\mathbf{R}|\xi\right)\right] = \left(\xi-1\right)\mathbb{E}_q\left[\log|\mathbf{R}|\right],$$

where $Constant$ does not depend on $\mathbf{V}$ or ν. Here, ψ_K is the multivariate digamma function.

Proof. First note that

$$\log|\mathbf{\Sigma}| = 2\log|\mathbf{S}| + \log|\mathbf{R}| = \sum_{k=1}^{K}\log\mathbf{S}_k^2 + \log|\mathbf{R}| = \sum_{k=1}^{K}\log\mathbf{\Sigma}_{kk} + \log|\mathbf{R}| \Rightarrow$$

$$\log|\mathbf{R}| = \log|\mathbf{\Sigma}| - \sum_{k=1}^{K}\log\mathbf{\Sigma}_{kk}.$$

By properties of the inverse Wishart distribution,

$$E_q\left[\log|\mathbf{\Sigma}|\right] = \log\left|\mathbf{V}^{-1}\right| - \psi_K\left(\frac{\nu}{2}\right) - K\log 2,$$

where ψ_p is the multivariate digamma function. By the marginalization property of the inverse Wishart distribution,

$$\mathbf{\Sigma}_{kk} \sim \text{InverseWishart}\left(\left(\mathbf{V}^{-1}\right)_{kk}, \nu - K + 1\right) \Rightarrow$$

$$E_q\left[\log\mathbf{\Sigma}_{kk}\right] = \log\left(\left(\mathbf{V}^{-1}\right)_{kk}\right) - \psi\left(\frac{\nu-K+1}{2}\right) - \log 2.$$

Plugging in gives the desired result. □

G Logistic GLMM Model Details

In this section we include extra details about the model and analysis of Section 4. We will continue to use the notation defined therein. Denoting by $Constant$ constants that do not depend on the prior parameters, parameters, or data, the log likelihood is

$$
\begin{aligned}
\log p\left(y_{it}|u_t,\beta\right) &= y_{it}\log\left(\frac{p_{it}}{1-p_{it}}\right)+\log\left(1-p_{it}\right)\\
&= y_{it}\rho+\log\left(1-p_{it}\right)+Constant\\
\log p\left(u|\mu,\tau\right) &= -\frac{1}{2}\tau\sum_{t=1}^{T}\left(u_t-\mu\right)^2-\frac{1}{2}T\log\tau\\
&= -\frac{1}{2}\tau\sum_{t=1}^{T}\left(u_t^2-\mu u_t+\mu^2\right)-\frac{1}{2}T\log\tau+Constant\\
\log p\left(\mu,\tau,\beta\right) &= -\frac{1}{2}\sigma_\mu^{-2}\left(\mu^2+2\mu\mu_0\right)+\\
&\quad \left(1-\alpha_\tau\right)\tau+\beta_\tau\log\tau+\\
&\quad -\frac{1}{2}\left(\text{trace}\left(\mathbf{\Sigma}_\beta^{-1}\beta\beta^T\right)+2\text{trace}\left(\mathbf{\Sigma}_\beta^{-1}\beta_0\beta^T\right)\right).
\end{aligned}
$$

The prior parameters were taken to be

$$
\begin{aligned}
\mu_0 &= 0.000\\
\sigma_\mu^{-2} &= 0.010\\
\beta_0 &= 0.000\\
\sigma_\beta^{-2} &= 0.100\\
\alpha_\tau &= 3.000\\
\beta_\tau &= 3.000.
\end{aligned}
$$

Under the variational approximation, ρ_{it} is normally distributed given x_{it}, with

$$
\begin{aligned}
\rho_{it} &= x_{it}^T\beta+u_t\\
\mathbb{E}_q\left[\rho_{it}\right] &= x_{it}^T\mathbb{E}_q\left[\beta\right]+\mathbb{E}_q\left[u_t\right]\\
\text{Var}_q\left(\rho_{it}\right) &= \mathbb{E}_q\left[\beta^T x_{it}x_{it}^T\beta\right]-\mathbb{E}_q\left[\beta\right]^T x_{it}x_{it}^T\mathbb{E}_q\left[\beta\right]+\text{Var}_q\left(u_t\right)\\
&= \mathbb{E}_q\left[\text{tr}\left(\beta^T x_{it}x_{it}^T\beta\right)\right]-\text{tr}\left(\mathbb{E}_q\left[\beta\right]^T x_{it}x_{it}^T\mathbb{E}_q\left[\beta\right]\right)+\text{Var}_q\left(u_t\right)\\
&= \text{tr}\left(x_{it}x_{it}^T\left(\mathbb{E}_q\left[\beta\beta^T\right]-\mathbb{E}_q\left[\beta\right]\mathbb{E}_q\left[\beta\right]^T\right)\right)+\text{Var}_q\left(u_t\right).
\end{aligned}
$$

We can thus use $n_{MC}=4$ points of Gauss-Hermite quadrature to numerically estimate $\mathbb{E}_q\left[\log\left(1-\frac{e^\rho}{1+e^\rho}\right)\right]$:

$$
\begin{aligned}
\rho_{it,s} &:= \sqrt{\text{Var}_q\left(\rho_{it}\right)}z_s+\mathbb{E}_q\left[\rho_{it}\right]\\
\mathbb{E}_q\left[\log\left(1-\frac{e^{\rho_{it}}}{1+e^{\rho_{it}}}\right)\right] &\approx \frac{1}{n_{MC}}\sum_{s=1}^{n_{MC}}\log\left(1-\frac{e^{\rho_{it,s}}}{1+e^{\rho_{it,s}}}\right)
\end{aligned}
$$

We found that increasing the number of points used for the quadrature did not measurably change any of the results. The integration points and weights were calculated using the `numpy.polynomial.hermite` module in python [Jones et al., 2001].