# Robust Inference with Variational Bayes

**Ryan Giordano**
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
rgiordano@berkeley.edu

Tamara Broderick
Department of EECS
Massachusetts Institute of Technology
Cambridge, MA 02139
tbroderick@csail.mit.edu

Michael Jordan
Department of EECS
University of California, Berkeley
Berkeley, CA 94720
jordan@cs.berkeley.edu

## 1 Introduction

In Bayesian analysis, the posterior follows from the data and a choice of a prior and a likelihood. One hopes that the posterior is robust to reasonable variation in the choice of prior and likelihood, since this choice is made by the modeler and is necessarily somewhat subjective. For example, the process of prior elicitation may be prohibitively time-consuming, two practitioners may have irreconcileable 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. If the posterior changes substantially with these choices of prior, then the analysis lacks objectivity. Measuring the sensitivity of the posterior to variation in the likelihood and prior is the central concern of the field of *robust Bayes*. A robust posterior is one that does not depend strongly on reasonable variation in the choice of model or prior, and robust Bayes provides methods for quantifying posterior robustness [1].

Despite the fundamental importance of the problem and a considerable body of literature, the tools of robust Bayes are not commonly used in practice. This is in large part due to the difficulty of calculating robustness measures from MCMC draws[2, 3]. Although methods for computing robustness measures from MCMC draws exist, they lack generality and often require additional coding or computation [1]. Consequently, formal robust Bayes methods are least used in complex, hierarchical models, exactly when they are needed most. Instead, modelers are tempted to either compute ad-hoc robustness estimates (e.g. by manually changing the priors and re-running their chain) or to ignore the problem altogether.

In contrast to MCMC, variational Bayes (VB) techniques are readily amenable to robustness analysis. The derivative of a posterior expectation with respect to a prior or data perturbation is a measure of *local robustness* to the prior or likelihood [4]. Because VB casts posterior inference as an optimization problem, its methodology is built on the ability to calculate derivatives of posterior quantities with respect to model parameters, even in very complex models. Variational methods for posterior approximation are increasingly providing a scalable alternative to MCMC for posterior

[1] See Appendix A for a literature review.

approximation, and this offers the opportunity to bring fast, easy-to-use robustness measures into common practice.

In the present work, we develop local prior robustness measures for *mean-field variational Bayes* (MFVB), a VB technique which imposes a particular factorization assumption on the variational posterior approximation. In past work [5], we demonstrated that a MFVB analysis can be quickly and straightforwardly augmented to provide information about local perturbations of the posterior variational approximation using linear response methods from statistical physics. We show that this framework can be extended to provide fast, easy-to-use prior robustness measures for posterior inference and thereby bring robustness analysis into common Bayesian practice.

In the remainder of the present work, we start by outlining existing local prior measures of robustness in Section 2. We extend the linear response techniques of [5] in Section 3. In Section 4 we use these results to derive closed-form measures of the sensitivity of mean-field variational posterior approximation to prior specification. In Section 5 we demonstrate our method on a meta-analysis of randomized controlled interventions in access to microcredit in developing countries.

## 2 Robustness measures

Denote our $N$ data points by $x = (x_1, \ldots, x_N)$ with $x_n \in \mathbb{R}^D$. Denote our parameter by the vector $\theta \in \mathbb{R}^K$. We denote the prior parameters by $\alpha$, where either $\alpha \in \mathbb{R}^M$ or $\alpha$ may be function-valued. Let $p_x^\alpha$ denote the posterior distribution of $\theta$, as given by Bayes' Theorem:

$$p_x^\alpha(\theta) := p(\theta|x,\alpha) = \frac{p(x|\theta)\, p(\theta|\alpha)}{p(x)}.$$

A typical end product of a Bayesian analysis might be a posterior expectation of some function $g(\theta)$ (e.g., a mean or variance): $\mathbb{E}_{p_x^\alpha}[g(\theta)]$, which is a functional of $g$. We suppose that we have determined that the prior parameter $\alpha$ belongs to some set $\mathcal{A}$, perhaps after expert prior elicitation. Finding the extrema of $\mathbb{E}_{p_x^\alpha}[g(\theta)]$ as $\alpha$ ranges over all of $\mathcal{A}$ is intractable or difficult except in special cases [6]. An alternative is to examine how much $\mathbb{E}_{p_x^\alpha}[g(\theta)]$ changes locally in response to small perturbations in the value of $\alpha$:

$$\left.\frac{d\mathbb{E}_{p_x^\alpha}[g(\theta)]}{d\alpha}\right|_\alpha \Delta\alpha \tag{1}$$

That is, we consider *local robustness* [4] properties in lieu of global ones. When $\alpha$ is function-valued, we take Eq. (1) to be a Gateaux derivative. By calculating Eq. (1) for all $\Delta\alpha \in \mathcal{A} - \alpha$, we can estimate the robustness of $\mathbb{E}_{p_x^\alpha}[g(\theta)]$ in a small neighborhood of $\alpha$.

## 3 Linear response variational Bayes and extensions

We next review and extend linear response perturbations to a mean-field variational Bayes posterior approximation [5] in order to quickly and easily evaluate Eq. (1). Let $q_x^\alpha$ denote the variational approximation to posterior $p_x^\alpha$. Recall that $q_x^\alpha$ is an approximate distribution selected to minimize the Kullback-Liebler divergence between $p_x^\alpha$ and $q$ across distributions $q$ in some class $\mathcal{Q}$. We consider the case where the variational family, $\mathcal{Q}$, is a class of products of exponential family distributions [7]:

$$\begin{aligned} q_x^\alpha &:= \operatorname{argmin}_{q\in\mathcal{Q}}\{S - L\} \quad \text{for} \quad \mathcal{Q} = \left\{ q : q(\theta) = \prod_{k=1}^K q(\theta_k); \quad \forall k, q(\theta_k) \propto \exp(\eta_k^T \theta_k) \right\} \\ L &:= \mathbb{E}_q[\log p(x|\theta)] + \mathbb{E}_q[\log p(\theta|\alpha)], \qquad S := \mathbb{E}_q[\log q(\theta)] \end{aligned} \tag{2}$$

We assume that $q_x^\alpha$, the solution to Eq. (2), has interior exponential family parameter $\eta_k$. In this case, $q_x^\alpha$ can be completely characterized by its mean parameters, $m := \mathbb{E}_{q_x^\alpha}[\theta]$ [8]. One can perturb the objective in Eq. (2) in the direction of a function $f$ of the mean parameter $m$ by some amount $t$, where $t$ is a vector with length equal to the output of $f$:

$$q_t \quad := \quad \mathrm{argmin}_{q \in \mathcal{Q}} \left\{ S - L + f(m)^T t \right\} \tag{3}$$

[5] showed that when $f(m) = m$, we can calculate the local change in the mean of $q_t$ as $t$ varies:

$$\left. \frac{d\mathbb{E}_{q_t}[\theta]}{dt^T} \right|_{t=0} = (I - VH)^{-1} V =: \hat{\Sigma}, \quad \text{where } V := \mathrm{Cov}_{q_x^\alpha}(\theta) \text{ and } H := \frac{\partial^2 L}{\partial m \partial m^T}. \tag{4}$$

As shown in Appendix B, if $f(m)$ and $h(m)$ are both smooth functions of $m$, then

$$\frac{dh(m_t)}{dt} = \nabla h^T \hat{\Sigma} \nabla f \tag{5}$$

Eq. (4) is the special case of Eq. (5) where $h(m) = f(m) = m$. In [5], the goal was to calculate a posterior covariance estimate $\hat{\Sigma}$. Here, our goal is to calculate a measure of robustness to changes in $\alpha$. Let $\alpha_t = \alpha + \Delta\alpha t$ be the value of $\alpha$ perturbed in direction $\Delta\alpha$ by an infinitesimal scalar amount $t$. $\Delta\alpha$ may be vector- or function-valued. Note that $\mathbb{E}_q[\log(p(\theta|\alpha))]$ from Eq. (2) is a function of $m$, since it is an expectation with respect to $q$, which is completely parameterized by $m$. Assuming that $p(\theta|\alpha)$ is a smooth function of $\alpha$, a Taylor expansion in $\Delta\alpha t$ gives

$$\begin{aligned} \mathbb{E}_q[\log(p(\theta|\alpha_t))] &= \mathbb{E}_q[\log(p(\theta|\alpha))] + \frac{d}{d\alpha^T}\mathbb{E}_q[\log(p(\theta|\alpha))]\Delta\alpha t + O(t^2) \Rightarrow \\ f(m) &:= \frac{d}{d\alpha^T}\mathbb{E}_q[\log(p(\theta|\alpha))]\Delta\alpha \quad \text{and} \quad h(m) := \mathbb{E}_{q_x^\alpha}[g(\theta)] \end{aligned} \tag{6}$$

With $f(m)$ and $h(m)$ defined as in Eq. (6), Eq. (5) gives us the robustness measure Eq. (1). As in LRVB, these derivatives are in fact the exact robustness of the variational posterior expectations to prior perturbation. The extent to which it represents the true prior sensitivity depends on the extent to which the MFVB means are good estimates of the true posterior means.

## 4 Robustness measures from LRVB

We now turn to calculating $f(m)$ from Eq. (6) for some common cases. For simplicity, we will take $g(\theta) = \theta$. First, consider a prior in the exponential family with sufficient statistics $\pi(\theta)$.

$$\log p(\theta|\alpha) \quad = \quad \alpha^T \pi(\theta) \Rightarrow f(m) = \mathbb{E}_{q_x^\alpha}[\pi(\theta)]\Delta\alpha \tag{7}$$

Here, $\pi(\theta)$ is a vector of the same length as $\alpha$. Note that $f(m)$ may be known exactly or estimated using Monte Carlo simulation. The simplest case is when the priors are conditionally conjugate for $p(x|\theta)$. In that case, $\pi(\theta) = \theta$, and $\frac{d\mathbb{E}_{q_x^\alpha}[\theta_i]}{d\alpha_j} = \hat{\Sigma}_{ij}$. A more complex non-conjugate example is the LKJ prior on a covariance matrix, which we explore in Section 5.

Next, we consider changing the functional form of $p(\theta|\alpha)$, taking $\Delta\alpha$ to be function-valued. We will focus on perturbations to the prior marginals, since local robustness properties of functional neighborhoods of the full posterior have bad asymptotic properties [9]. Let $\theta_i$ be a subvector of $\theta$ whose marginal we will perturb. We assume that both the prior and variational distribution factor across $\theta_i$:

$$q_x^\alpha(\theta) = q(\theta_i)q(\theta_{-i}) \quad \text{and} \quad p(\theta|\alpha) = p(\theta_i|\alpha_i)p(\theta_{-i}|\alpha_{-i})$$

where $-i$ denotes $1, ..., K \setminus i$. For simplicity of notation, assume without loss of generality that the $i$ indices come first: $\theta^T = (\theta_i^T, \theta_{-i}^T)$ (Both $q_x^\alpha$ and the prior may factorize still further.)

In order to ensure that the perturbed prior is properly normalized, we will shift an infinitesimal amount of prior mass from the original $p(\theta_i|\alpha)$ to a density $p_c(\theta_i)$:

$$p(\theta_i|\alpha_i, \epsilon) = (1 - \epsilon)p(\theta_i|\alpha_i) + \epsilon p_c(\theta_i) \tag{8}$$

This is known as $\epsilon$-contamination, and its construction guarantees that the perturbed prior is properly normalized [2]. By taking $p_c(\theta_i) = \delta(\theta_i - \theta_{i0})$ to be a Dirac delta function at $\theta_{i0}$, Eq. (5) and Eq. (6) give (see Appendix C):

$$\frac{d\mathbb{E}_q[\theta]}{d\epsilon} = \frac{q_x^\alpha(\theta_{i0})}{p(\theta_{i0}|\alpha)}(I - VH)^{-1}\begin{pmatrix} \theta_{i0} - m_i \\ 0 \end{pmatrix} \tag{9}$$

This is known as an "influence function" [4]. Note that $p(\theta_{i0}|\alpha)$ is known *a priori*, and that $q_x^\alpha(\theta_{i0})$ is a function of the moment parameters $m$, since $m$ entirely specifies $q_x^\alpha$. Viewed as a function of $\theta_0$, Eq. (9) characterizes how much each moment parameter, $m$, is affected by adding an infinitesimal amount of prior mass at $\theta_{i0}$. By the linearity of the derivative, one can use weighted combinations of delta functions and Eq. (9) to estimate the sensitivity to any prior function [3].

## 5 Experiments

We applied the methods above to a hierarchical model of microcredit interventions in development economics [11]. One output of the model is $\mu$ and $\tau$, top level parameters in a hierarchical model that measure average site profitability and the effectiveness of microcredit interventions, respectively. Here, we present the sensitivity of these parameters to $\Lambda$, the information matrix of a normal prior on $(\mu, \tau)$, and $\eta$, the concentration parameter in a non-conjugate LKJ prior[12] on the covariance of $(\mu, \tau)$. The left panel of Fig. (1) shows the estimates from Eq. (7) normalized by the posterior standard deviation. The results are robust to $\eta$ but extremely non-robust to $\Lambda$. The second panel compares the prediction of Eq. (7) to the actual change in MCMC means to a small change in $\Lambda_{11}$. The results match closely. The third panel shows Eq. (9), the influence function of the prior for $(\mu, \tau)$ on $\tau$. The "X" is the posterior mean. Adding prior mass on only one side of the mean would be highly influential, though it is hard to imagine such a prior representing an *a priori* belief.

We formed the LRVB estimates using JuMP[13] and used STAN[14] to generate MCMC samples. The VB and MCMC results are nearly identical, indicating that the assumptions necessary for LRVB hold. Generating one set of MCMC draws took 15 minutes, and the LRVB estimates, including calculating all the reported sensitivity measures, took 45 seconds. For more details, see Appendix D.

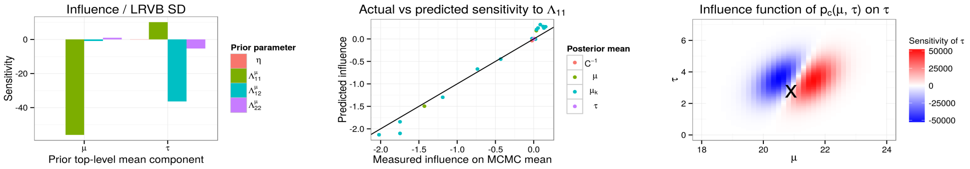

Figure 1: Effectiveness of robustness measures in the microcredit model

[2] $\epsilon$-contamination is principally adopted for analytic convenience, though it is an expressive class of perturbations [9]. For more exotic perturbation classes, which we do not consider here, see[10].

[3] A closed form for $p_c(\theta_i)$ other than weighted combinations of Dirac delta functions is given in Appendix C.2. The influence function is closely related to the worst-case prior perturbation within a metric ball in the space of prior functions [9]. We show in Appendix C.3 that LRVB also gives a closed form for this worst-case perturbation. Appendix C.4 provides some intuition by comparing the LRVB results to the corresponding formulas for exact inference.

# References

[1] David Ríos Insua and Fabrizio Ruggeri. *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2012.

[2] James O. Berger, David Ríos Insua, and Fabrizio Ruggeri. Robust bayesian analysis. In David Ríos Insua and Fabrizio Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2012.

[3] Małgorzata Roos, Thiago G Martins, Leonhard Held, Håvard Rue, et al. Sensitivity analysis for bayesian hierarchical models. *Bayesian Analysis*, 10(2):321–349, 2015.

[4] Paul Gustafson. Local robustness in bayesian analysis. In David Ríos Insua and Fabrizio Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2012.

[5] Ryan Giordano, Tamara Broderick, and Michael Jordan. Linear response methods for accurate covariance estimates from mean field variational bayes. *arXiv preprint arXiv:1506.04088*, 2015.

[6] Elias Moreno. Global bayesian robustness for some classes of prior distributions. In David Ríos Insua and Fabrizio Ruggeri, editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media, 2012.

[7] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. Chapter 10.

[8] 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.

[9] Paul Gustafson et al. Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195, 1996.

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

[11] Rachael Meager. Understanding the impact of microcredit expansions: A bayesian hierarchical analysis of 7 randomised experiments. *arXiv preprint arXiv:1506.06669*, 2015.

[12] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001, 2009.

[13] M. Lubin and I. Dunning. Computing in operations research using Julia. *INFORMS Journal on Computing*, 27(2):238–248, 2015.

[14] *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*, 2015.

[15] Paul Gustafson. Local sensitivity of inferences to prior marginals. *Journal of the American Statistical Association*, 91(434):774–781, 1996.

[16] CJ Pérez, J Martín, and MJ Rufo. Mcmc-based local parametric sensitivity estimations. *Computational Statistics & Data Analysis*, 51(2):823–835, 2006.

[17] CJ Pérez, J Martín, and MJ Rufo. Sensitivity estimations for bayesian inference models solved by mcmc methods. *Reliability Engineering & System Safety*, 91(10):1310–1314, 2006.

[18] Robbert E Kass, Luke Tierney, and Joseph B Kadane. Approximate methods for assessing influence and sensitivity in bayesian analysis. *Biometrika*, 76(4):663–674, 1989.

[19] Robert E McCulloch. Local model influence. *Journal of the American Statistical Association*, 84(406):473–478, 1989.

[20] Luke Bornn, Arnaud Doucet, and Raphael Gottardo. An efficient computational approach for prior sensitivity analysis and cross-validation. *Canadian Journal of Statistics*, 38(1):47–64, 2010.

[21] Donald B Rubin. Estimation in parallel randomized experiments. *Journal of Educational and Behavioral Statistics*, 6(4):377–401, 1981.

[22] John Barnard, Robert McCulloch, and Xiao-Li Meng. Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Statistica Sinica*, 10(4):1281–1312, 2000.

# Appendices

## A Robust Bayes with MCMC

There is an extensive literature on Robust Bayesian techniques, surveyed in [1]. We focus on local robustness techniques [4, 9, 15]. In the original papers, many authors focused on either theoretical results or models with special structure that rendered robustness measures tractable. Of MCMC, one of the founders of the field of Bayesian Robustness writes:

"The MCMC methodology was not directly compatible with many of the robust Bayesian techniques that had been developed, so that it was unclear how formal robust Bayesian analysis could be incorporated into the future 'Bayesian via MCMC' world. Paradoxically, MCMC has dramatically increased the need for consideration of Bayesian robustness, in that the modeling that is now routinely utilized in Bayesian analysis is of such complexity that inputs (such as priors) can be elicited only in a very casual fashion."[2]

Another recent author adds:

"Surprisingly, despite considerable theoretical advances in formal sensitivity analysis, it is barely used in every-day practice... a formal robustness methodology which is feasible, fairly quick, operating with low extra computing effort and provided by default in a dedicated software, is strongly required." [3]

A number of papers have proposed methods for performing robustness analyses using MCMC techniques. [15], following many previous theoretical works [4], exchanges the integral in a posterior expectation with the derivative with respect to prior perturbations, giving a robustness estimate that can be evaluated from MCMC samples. [16, 17] extends this idea. These approaches exploit importance sampling and / or closed forms for derivatives or posterior densities, and care must be taken to control the variance of the MCMC estimates. The papers [18, 19] make second-order approximations to the log posterior and employ numerical techniques to calculate robustness measures. [3] uses a sophisticated methodology to choose a grid of prior points at which they numerically estimate the sensitivity using estimates of the posterior density. [20] proposes a distinctive method based on particle filtering in which particle weights are re-adjusted to produce draws from a perturbed prior. The authors are unaware of any previous work applying robust Bayes techniques in the context of variational methods.

The advantage of using robust Bayes with LRVB over these MCMC-based techniques is simplicity and computational ease. Little extra code and no extra approximations or assumptions beyond that required for LRVB are required to compute the robustness measures below. LRVB robustness measures are the exact sensitivity of the variational solution to changes in the prior, and they will be accurate to the extent that the variational approximation to the posterior mean of interest is accurate [5].

## B LRVB covariance of functions

Let us consider LRVB estimates of the covariances of functions of natural parameters rather than the natural parameters themselves. Suppose we have a function $\phi\left(\eta\right)$, and a variational solution $q\left(m\right)$ where $m=\mathbb{E}_{q}\left[\eta\right]$. Since $q$ is fully parameterized by $m$, we can write

$$\mathbb{E}_{q}\left[\phi\left(\eta\right)\right] = f\left(m\right)$$

for some continuous $f\left(m\right)$. We can consider a perturbed log likelihood that also includes $f\left(m\right)$:

$$\begin{aligned}
\log p_t &= \log p + t_0^T m + t_f f\left(m\right) := \log p + t^T m_f \\
t &:= \left(\begin{array}{c} t_0 \\ t_f \end{array}\right) \\
m_f &:= \left(\begin{array}{c} m \\ f\left(m\right) \end{array}\right)
\end{aligned}$$

As in [5], we use the fixed point equations:

$$\begin{aligned}
E_t &:= E + t^T m_f \\
\frac{dE_t}{dm} &= 0 \Rightarrow \\
\frac{dE}{dm} + \left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right) &= 0 \\
M\left(m\right) &:= \frac{\partial E}{\partial m} + m \\
M_t\left(m\right) &:= M\left(m\right) + \left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right) \\
M_t\left(m^*\right) &:= m^* \text{ (definition of } m^*\text{)} \\
\frac{dm_t^*}{dt^T} &= \left.\frac{\partial M_t}{\partial m^T}\right|_{m=m_t^*} \frac{dm_t^*}{dt^T} + \frac{\partial M_t}{\partial t^T} \\
&= \left(\left.\frac{\partial M}{\partial m^T}\right|_{m=m_t^*} + \frac{\partial}{\partial m^T}\left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right)\right)\frac{dm^*}{dt^T} + \left(\begin{array}{cc} I & \nabla f \end{array}\right)
\end{aligned}$$

The term $\frac{\partial}{\partial m^T}\left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right)$ is awkward, but it disappears when we evaluate at $t = 0$, giving

$$\begin{aligned}
\frac{dm_t^*}{dt^T} &= \left(\left.\frac{\partial M}{\partial m^T}\right|_{m=m_t^*}\right)\frac{dm^*}{dt^T} + \left(\begin{array}{cc} I & \nabla f \end{array}\right) \\
&= \left(\frac{\partial^2 E}{\partial m \partial m^T} + I\right)\frac{dm^*}{dt^T} + \left(\begin{array}{cc} I & \nabla f \end{array}\right) \\
\frac{dm^*}{dt^T} &= -\left(\frac{\partial^2 E}{\partial m \partial m^T}\right)^{-1}\left(\begin{array}{cc} I & \nabla f \end{array}\right)
\end{aligned}$$

Recalling that

$$\frac{dm^*}{dt_0^T} := \hat{\Sigma}$$

We can plug in to see that

$$\frac{dm^*}{dt_f^T} = \hat{\Sigma}\nabla f$$

This means that the covariance of the natural sufficient statistics with the function $\phi(\eta)$ are determined by a linear combination of the LRVB covariance matrix.

A similar conclusion can be reached by considering the response of the expectation of a quantity other than a natural parameter to a generic perturbation. Consider perturbing the log likelihood by some function $t_g g(m)$. Then by the reasoning above,

$$\frac{df(m)}{dt_g} = \frac{df}{dm^T}\frac{dm}{dt_g} = \nabla f^T \hat{\Sigma} \nabla g$$

This is Eq. (5), and represents the LRVB covariance between two quantities with variational expectation $f(m)$ and $g(m)$ respectively. As in the present, that covariance can also be interpreted as the sensitivity of $g(m)$ to a perturbation of the objective by $g(m)$.

# C Robustness Derivations

In this section, we derive results stated in Section 4. For generality, when possible we will derive results for the full vector $\theta$ rather than the sub-vector $\theta_i$ when the proof would be identical for the subvector under the assumption that $q(\theta) = q(\theta_i)q(\theta_{-i})$.

## C.1 Sensitivity to $\epsilon-$contamination

For a given $p_c(\theta)$ in Eq. (8), we can consider $p(\theta|\alpha, \epsilon)$ to be a class of priors parameterized by $(\alpha, \epsilon)$, and take $\epsilon = \Delta\alpha$ in Eq. (6). We then need to calculate

$$\begin{aligned} \frac{d}{d\epsilon}\mathbb{E}_q\left[\log p(\theta|\alpha,\epsilon)\right]\Big|_{\epsilon=0} &= \mathbb{E}_q\left[\frac{d}{d\epsilon}\log\left((1-\epsilon)p(\theta|\alpha)+\epsilon p_c(\theta)\right)\Big|_{\epsilon=0}\right] \\ &= \mathbb{E}_q\left[\frac{p_c(\theta)}{p(\theta|\alpha)} - 1\right] \end{aligned}$$

Since the variational solution is unaffected by adding constants to the ELBO, we can take

$$f(m) := \mathbb{E}_q\left[\frac{p_c(\theta)}{p(\theta|\alpha)}\right] \tag{10}$$

## C.2 Sensitivity to a function

We will calculate $\nabla f(m)$ using Eq. (10) for a general function $p_c(\theta)$ and then use this result to derive Eq. (9) as a special case. In this section, we rely on the fact that the variational distribution, $q(\theta)$, is in the exponential family.

The directional derivative for a perturbation $p_c(\theta)$ is given by the Taylor expansion of $\mathbb{E}_q\left[\frac{p_c(\theta)}{p(\theta|\alpha)}\right]$ in terms of the exponential family moment parameters:

$$\begin{aligned} \frac{d}{dm}\mathbb{E}_q\left[\frac{p_c(\theta)}{p(\theta|\alpha)}\right] &= V^{-1}\frac{d}{d\eta}\int \exp\left(\eta^T\theta - A(\theta)\right)\frac{p_c(\theta)}{p(\theta|\alpha)}d\theta \\ &= V^{-1}\int q(\theta)(\theta - m)\frac{p_c(\theta)}{p(\theta|\alpha)}d\theta \\ &= V^{-1}\mathbb{E}_q\left[(\theta - m)\frac{p_c(\theta)}{p(\theta|\alpha)}\right] \end{aligned} \tag{11}$$

Taking $p_c(\theta) = \delta\left(\theta_i = \theta_{i0}\right)$ to be a Dirac delta function gives Eq. (9).

### C.3 Extremal derivative

The influence function is closely related to the worst-case prior perturbation in a metric ball around the original prior, $p(\theta_i|\alpha_i)$. We refer the reader to [9] for the background. Given Eq. (11), the proof for the variational case is essentially identical.

First, to match [9], let $p_c(\theta)$ be a signed measure and consider perturbations of the form

$$p(\theta|\alpha, \epsilon) = p(\theta|\alpha) + \epsilon p_c(\theta)$$

Because the variational solution is invariant to constants, the variational sensitivity to this perturbation is identical to that of $\epsilon-$ contamination. Consequently, the sensitivity is given by Eq. (11) and Eq. (5):

$$\begin{aligned} \frac{d\mathbb{E}_q\left[g(\theta)\right]}{dt} &= \nabla h^T \hat{\Sigma} V^{-1} \mathbb{E}_q\left[\left(\theta - m\right)\frac{p_c(\theta)}{p\left(\theta|\alpha\right)}\right] \\ &= \mathbb{E}_q\left[\nabla h^T \left(I - VH\right)^{-1}\left(\theta - m\right)\frac{p_c(\theta)}{p\left(\theta|\alpha\right)}\right] \end{aligned}$$

Define

$$a\left(\theta\right) = \nabla h^T \left(I - VH\right)^{-1}\left(\theta - m\right)\frac{q\left(\theta\right)}{p\left(\theta|\alpha\right)}$$

As in [9], for $p \in [1, \infty]$ and $\frac{1}{p} + \frac{1}{q} = 1$, define the size of a perturbation as

$$\left(\int \left|\frac{p_c(\theta)}{p(\theta|\alpha)}\right|^p d\Pi\right)^{\frac{1}{p}} \tag{12}$$

...where $\Pi$ is the measure on $\theta$ induced by $p(\theta|\alpha)$ and $p \in [1, \infty]$. Let $(\cdot)^+$ denote the positive part and $(\cdot)^-$ the negative part of the term in the parentheses.

$$\begin{aligned} \mathbb{E}_q\left[\left|R^T\left(I - VH\right)^{-1}\left(\theta - m\right)\frac{q\left(\theta\right)}{p\left(\theta|\alpha\right)}\right|^+\right] &= \int\left|a\left(\theta\right)^+\frac{p_c(\theta)}{p\left(\theta|\alpha\right)}\right| d\Pi \\ &\le \left(\int\left|a\left(\theta\right)^+\right|^q d\Pi\right)^{\frac{1}{q}}\left(\int\left|\frac{p_c(\theta)}{p\left(\theta|\alpha\right)}\right|^p d\Pi\right)^{\frac{1}{p}} \\ &= \left(\int\left|a\left(\theta\right)^+\right|^q d\Pi\right)^{\frac{1}{q}} \end{aligned}$$

Since we are taking $p_c(\theta)$ such that $\|\frac{p_c(\theta)}{p(\theta|\alpha)}; \Pi\|_p = 1$. This is maximized when

$$\begin{aligned} \left|a\left(\theta\right)^+\right|^q &\propto \left|\frac{p_c(\theta)}{p\left(\theta|\alpha\right)}\right|^p \\ p_c(\theta) &= \pi\left|\left(R^T\left(I - VH\right)^{-1}\left(\theta - m\right)\right)^+\frac{q\left(\theta\right)}{p\left(\theta|\alpha\right)}\right|^{\frac{q}{p}} \end{aligned}$$

A similar analysis follows for $a(\theta)^-$, and it follows that the worst-case prior perturbation in a $p-$neighborhood of $p(\theta|\alpha)$ is given by

$$p_c(\theta) \quad = \quad p(\theta|\alpha) \max \left\{ \left|a(\theta)^+\right|^{\frac{1}{p-1}}, \left|a(\theta)^-\right|^{\frac{1}{p-1}} \right\} \tag{13}$$

### C.4 Comparison with Exact Results

Comparing Eq. (13) with [9, Equation 6] lends some intuition. In our notation, the exact extremal perturbation is given by Eq. (14) by the same expression as Eq. (13) but with a different $a(\theta)$:

$$a_p(\theta) \quad = \quad g(\theta) \left(\theta - \mathbb{E}_{p_x^\alpha}[\theta]\right) \frac{p(\theta|x)}{p(\theta|\alpha)} \tag{14}$$

Here, $q_x^\alpha$ plays the role of the marginal posterior $p(\theta|x)$, and $\mathbb{E}_{q_x^\alpha}[g(\theta)]^T (I - VH)^{-1}$ plays the role of $g(\theta)$. Note that a principal difficulty of using Eq. (14) is that Eq. (14) requires knowledge of ratio of the posterior density to the prior density, which is not automatically available from MCMC draws. The MFVB solution circumvents this difficulty by providing an explicit parametric approximation to the posterior density.

## D Microcredit Model

We will reproduce a variant of the analysis performed in [11], though with somewhat different prior choices. Randomized controlled trials were run in seven different sites to try to measure the effect of access to microcredit on various measures of business success. Each trial was found to lack power individually for various reasons, so there could be some benefit to pooling the results in a simple hierarchical model. For the purposes of demonstrating robust Bayes techniques with VB, we will focus on the simpler of the two models in [11] and ignore covariate information.

We will index sites with $k = 1, .., K$ (here, $K = 7$) and business within a site by $i = 1, ..., N_k$. In site $k$ and business $i$ we observe whether the business was randomly selected for increased access to microcredit, denoted $T_{ik}$, and the profit after intervention, $y_{ik}$. We follow [21] and assume that each site has an idiosyncratic average profit, $\mu_k$ and average improvement in profit, $\tau_k$, due to the intervention. Given $\mu_k$, $\tau_k$, and $T_{ik}$, the observed profit is assumed to be generated according to

$$y_{it}|\mu_k, \tau_k, T_{ik}, \sigma_k \quad \sim \quad N\left(\mu_k + T_{ik}\tau_k, \sigma_k^2\right)$$

The site effects, $(\mu_k, \tau_k)$, are assumed to come from an overall pool of effects and may be correlated:

$$\begin{aligned} \begin{pmatrix} \mu_k \\ \tau_k \end{pmatrix} \quad &\sim \quad N\left(\begin{pmatrix} \mu \\ \tau \end{pmatrix}, C\right) \\ C \quad &:= \quad \begin{pmatrix} \sigma_\mu^2 & \sigma_{\mu\tau} \\ \sigma_{\mu\tau} & \sigma_\tau^2 \end{pmatrix} \end{aligned}$$

The effects $\mu$, $\tau$, and the covariance matrix $V$ are unknown parameters that require priors. For $(\mu, \tau)$ we simply use a bivariate normal prior. However, choosing an appropriate prior for a covariance matrix can be conceptually difficult [22]. Following the recommended practice of the software package STAN[14], we derive a variational model to accommodate the non-conjugate LKJ prior [12], allowing the user to model the covariance and marginal variances separately. Specifically, we use

$$\begin{aligned} C &=: SRS \\ S &= \text{Diagonal matrix} \\ R &= \text{Covariance matrix} \\ S_{kk} &= \sqrt{\text{diag}(C)_k} \end{aligned}$$

We can then put independent priors on the scale of the variances, $S_{kk}$, and on the covariance matrix, $R$. We model the inverse of $C$ with a Wishart variational distribution, and use the following priors:

$$\begin{aligned} q\left(C^{-1}\right) &= \text{Wishart}(V_{\Lambda}, n) \\ p\left(S\right) &= \prod_{k=1}^{2} p(S_{kk}) \\ S_{kk}^2 &\sim \text{InverseGamma}(\alpha_{scale}, \beta_{scale}) \\ \log p(R) &= (\eta - 1)\log|R| + C \end{aligned}$$

The necessary expectations have closed forms with the Wishart variational approximation, as derived in Appendix E.

In addition, we put a normal prior on $(\mu, \tau)^T$ and an inverse gamma prior on $\sigma_k^2$:

$$\begin{aligned} \left(\begin{array}{c} \mu \\ \tau \end{array}\right) &\sim N\left(\left(\begin{array}{c} 0 \\ 0 \end{array}\right), \Lambda^{-1}\right) \\ \sigma_k^2 &\sim \text{InverseGamma}(\alpha_\tau, \beta_\tau) \end{aligned}$$

The prior parameters used were:

$$\begin{aligned} \Lambda &= \left(\begin{array}{cc} 0.02 & 0 \\ 0 & 0.02 \end{array}\right) \\ \eta &= 15.01 \\ \sigma_k^{-2} &\sim \text{InverseGamma}(2.01, 2.01) \\ \alpha_{scale} &= 20.01 \\ \beta_{scale} &= 20.01 \\ \alpha_\tau &= 2.01 \\ \beta_\tau &= 2.01 \end{aligned}$$

### D.1 Results

First, note that the the MCMC results match the VB means very closely, indicating that the assumptions underlying LRVB are satisfied. The least- well estimated parameters are $C^{-1}$.

We will focus on the robustness of $\mu$ and $\tau$, since as the higher-level parameters in the hierarchical model, they are both more susceptible to prior influence and more generally interpretable (as the

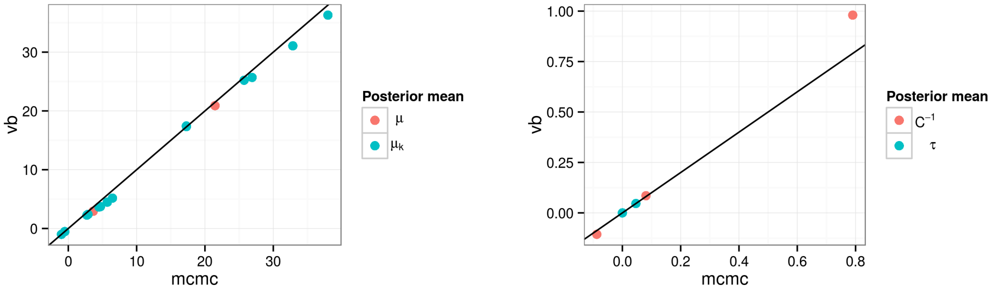

Figure 2: Comparison of MCMC and VB Results for the microcredit data

average profit and the causal effect of microcredit, respectively). The sensitivity of $(\mu, \tau)$ to $\Lambda$ and $\eta$ is shown in the left panel of Fig. (3) as a proportion of the LRVB posterior standard deviation. The parameters can be seen to be quite sensitive to changes in $\Lambda$. For example, if the upper left component of $\Lambda$, $\Lambda_{11}$, were to increase by $0.04$, $\mathbb{E}_q[\mu]$ would be expected to increase by two posterior standard deviations. If $0.06$ is a subjectively reasonable value for $\Lambda_{11}$, then the ordinary posterior confidence interval for $\mu$ is quite inadequate in capturing the subjective range of beliefs that might be assigned to $\mu$. In contrast, the sensitivty to $\eta$, the LKJ parameter, is quite small.

The right panel of Fig. (3) shows the influence function of $(\mu, \tau)$ on $\tau$. The $X$ marks the posterior mean. Recall that the prior mean is $(0, 0)$ and relatively diffuse. The numbers are quite large, indicating that adding a small amount of prior mass precisely near the posterior could influence the posterior considerably. However, such a prior perturbation would have to have informed by the data – adding mass nearly anywhere else would have a much smaller effect. What kind of prior perturbation is reasonable remains a subjective decision of the modeler.

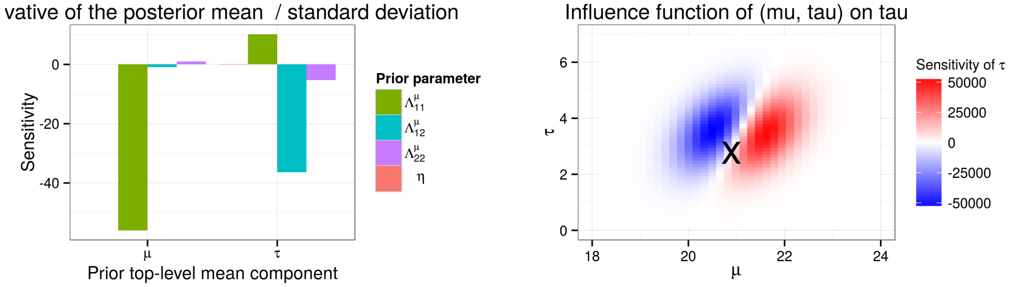

Figure 3: The sensitivity of $\mu$ and $\tau$

Finally, Fig. (4) shows the effects of changing $\Lambda_{11}$ on a re-run MCMC chain compared with the effects predicted by LRVB robustness measurements. The results are very good for all except $C^{-1}$, which was not estimated well by the VB model. Even for $C^{-1}$, the LRVB estimates are directionally correct.

## E 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.

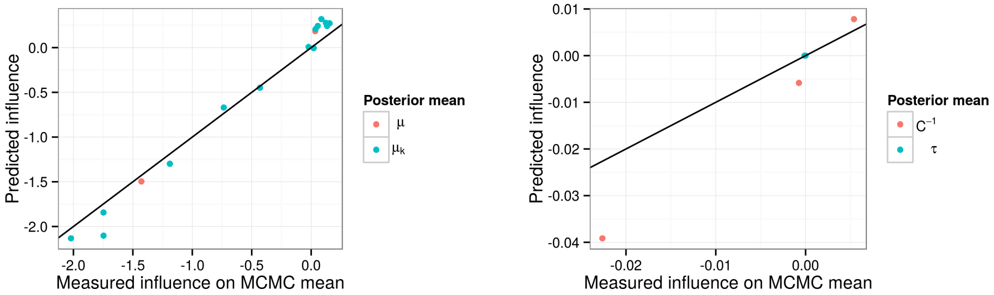

Figure 4: Predicted vs Actual effects of perturbations

We want to estimate a multivariate normal covariance matrix with flexible priors. For simplicity, let us study in isolation the model:

$$
\begin{aligned}
\log p\left(y|\Lambda\right) &= -\frac{1}{2}y^{T}\Lambda y+\frac{1}{2}\log\Lambda \\
\Lambda &= \Sigma^{-1} \\
\Sigma &=: SRS \\
S_{k} &:= \sqrt{diag\left(\Sigma\right)_{k}} \\
\log p\left(S\right) &= \sum_{k=1}^{K}\log p\left(S_{k}\right) \\
\log p\left(R\right) &= \log\left(C\left|R\right|^{\eta-1}\right) \\
&= \left(\eta-1\right)\log\left|R\right|+C \\
&= \text{ (LKJ prior)}
\end{aligned}
$$

Let us use a Wishart variational distribution for $\Lambda$:

$$
\begin{aligned}
q\left(\Lambda\right) &= \text{Wishart}\left(V,n\right) \\
E_{q}\left[\Lambda\right] &= nV \\
E_{q}\left[\log\left|\Lambda\right|\right] &= \psi_{p}\left(\frac{n}{2}\right)+\log\left|V\right|+K\log2 \\
\psi_{p}\left(n\right) &= \sum_{i=1}^{p}\psi\left(\frac{2n+1-i}{2}\right)
\end{aligned}
$$

Then $\Sigma$ has an inverse Wishart distribution:

$$\begin{aligned}
E_q\left[\Sigma\right] &= \frac{V^{-1}}{n-K-1} \\
\Sigma_{kk} &\sim \text{InverseWishart}\left(\left(V^{-1}\right)_{kk}, n-K+1\right) \\
E_q\left[\Sigma_{kk}\right] &= \frac{\left(V^{-1}\right)_{kk}}{n-K+1-2} = \frac{\left(V^{-1}\right)_{kk}}{n-K-1} \\
\log p\left(\Sigma_{kk}\right) &= -\left(\frac{(n-K+1)+1+1}{2}\right)\log\Sigma_{kk} - \frac{1}{2}\frac{\left(V^{-1}\right)_{kk}}{\Sigma_{kk}} + C \\
&= \left(-\frac{n-K+1}{2} - 1\right)\log\Sigma_{kk} - \frac{\frac{1}{2}\left(V^{-1}\right)_{kk}}{\Sigma_{kk}} + C \\
&= \log\left(\text{InvGamma}\left(\frac{n-K+1}{2}, \frac{1}{2}\left(V^{-1}\right)_{kk}\right)\right) \Rightarrow \\
E_q\left[\log\Sigma_{kk}\right] &= \log\left(\frac{1}{2}\left(V^{-1}\right)_{kk}\right) - \psi\left(\frac{n-K+1}{2}\right)
\end{aligned}$$

We'll also need the expectation of the square root of an inverse gamma distributed variable.

$$\begin{aligned}
p\left(x\right) &= \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{-\alpha-1}\exp\left(\frac{-\beta}{x}\right) \\
E\left[x^{\frac{1}{2}}\right] &= \int\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}x^{-\alpha-1+\frac{1}{2}}\exp\left(\frac{-\beta}{x}\right)dx \\
&= \int\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\frac{\beta^{\alpha-\frac{1}{2}}}{\Gamma\left(\alpha-\frac{1}{2}\right)}\frac{\Gamma\left(\alpha-\frac{1}{2}\right)}{\beta^{\alpha-\frac{1}{2}}}x^{-\left(\alpha-\frac{1}{2}\right)-1}\exp\left(\frac{-\beta}{x}\right)dx \\
&= \frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}\frac{\Gamma\left(\alpha-\frac{1}{2}\right)}{\beta^{\alpha-\frac{1}{2}}} \\
&= \beta^{\frac{1}{2}}\frac{\Gamma\left(\alpha-\frac{1}{2}\right)}{\Gamma\left(\alpha\right)}
\end{aligned}$$

Thus

$$E_q\left[\sqrt{\Sigma_{kk}}\right] = E_q\left[S_k\right] = \sqrt{\frac{1}{2}\left(V^{-1}\right)_{kk}}\frac{\Gamma\left(\frac{n-K}{2}\right)}{\Gamma\left(\frac{n-K+1}{2}\right)}$$

This means we have a closed form expectation of the LKJ prior. For the scale parameters, we can use a gamma prior distribution:

$$\begin{aligned}
\log p\left(S_k\right) &= \log\Gamma\left(\alpha,\beta\right) \\
&= -\beta S_k + \left(\alpha-1\right)\log S_k + C \\
&= -\beta S_k + \frac{\left(\alpha-1\right)}{2}\log S_k^2 + C
\end{aligned}$$

Finally, these expectations are given in terms of the natural parameters, but for LRVB we need derivatives with respect to the mean parameters. In the Wishart distribution, the mapping from mean parameters to natural parameters does not have a closed form. Eq. (4) requires the derivatives of the likelihood with respect to the moment parameters, and the Hessian must be transformed before use. Note that the Hessian of the likelihood is not necessarily at a maximum, so the transform requires a third-order tensor product.