Linear Response Methods for Accurate Covariance Estimates from Mean Field Variational Bayes

Ryan Giordano
UC Berkeley
rgiordano@berkeley.edu

Tamara Broderick
MIT
tbroderick@csail.mit.edu

Michael Jordan
UC Berkeley
jordan@cs.berkeley.edu

December 8, 2015

Abstract

Mean field variational Bayes (MFVB) is a popular posterior approximation method due to its fast runtime on large-scale data sets. However, a well known major failing of MFVB is that it underestimates the uncertainty of model variables (sometimes severely) and provides no information about model variable covariance. We generalize linear response methods from statistical physics to deliver accurate uncertainty estimates for model variables—both for individual variables and coherently across variables. We call our method *linear response variational Bayes* (LRVB). When the MFVB posterior approximation is in the exponential family, LRVB has a simple, analytic form, even for non-conjugate models. Indeed, we make no assumptions about the form of the true posterior. We demonstrate the accuracy and scalability of our method on a range of models for both simulated and real data.

1 Introduction

With increasingly efficient data collection methods, scientists are interested in quickly analyzing ever larger data sets. In particular, the promise of these large data sets is not simply to fit old models but instead to learn more nuanced patterns from data than has been possible in the past. In theory, the Bayesian paradigm yields exactly these desiderata. Hierarchical modeling allows practitioners to capture complex relationships between variables of interest. Moreover, Bayesian analysis allows practitioners to quantify the uncertainty in any model estimates—and to do so coherently across all of the model variables.

Mean field variational Bayes (MFVB), a method for approximating a Bayesian posterior distribution, has grown in popularity due to its fast runtime on large-scale data sets [3, 4, 6]. But a well known major failing of MFVB is that it gives underestimates of the uncertainty of model variables that can be arbitrarily bad, even when approximating a simple multivariate Gaussian distribution [2, 11, 20]. Also, MFVB provides no information about how the uncertainties in different model variables interact [2, 17, 20, 23].

By generalizing linear response methods from statistical physics [13–15, 19] to exponential family variational posteriors, we develop a methodology that augments MFVB to deliver accurate uncer-

tainty estimates for model variables—both for individual variables and coherently across variables. In particular, as we elaborate in Section 2, when the approximating posterior in MFVB is in the exponential family, MFVB defines a fixed-point equation in the means of the approximating posterior, and our approach yields a covariance estimate by perturbing this fixed point. We call our method *linear response variational Bayes* (LRVB).

We provide a simple, intuitive formula for calculating the linear response correction by solving a linear system based on the MFVB solution (Section 2.2). We show how the sparsity of this system for many common statistical models may be exploited for scalable computation (Section 2.3). We demonstrate the wide applicability of LRVB by working through a diverse set of models to show that the LRVB covariance estimates are nearly identical to those produced by a Markov Chain Monte Carlo (MCMC) sampler, even when MFVB variance is dramatically underestimated (Section 3). Finally, we focus in more depth on models for finite mixtures of multivariate Gaussians (Section 3.3), which have historically been a sticking point for MFVB covariance estimates [2, 20]. We show that LRVB can give accurate covariance estimates orders of magnitude faster than MCMC (Section 3.3). We demonstrate both theoretically and empirically that, for this Gaussian mixture model, LRVB scales linearly in the number of data points and approximately cubically in the dimension of the parameter space (Section 3.4).

Previous Work. Linear response methods originated in the statistical physics literature [8, 13, 14, 19]. These methods have been applied to find new learning algorithms for Boltzmann machines [8], covariance estimates for discrete factor graphs [24], and independent component analysis [7]. [18] states that linear response methods could be applied to general exponential family models but works out details only for Boltzmann machines. [14], which is closest in spirit to the present work, derives general linear response corrections to variational approximations; indeed, the authors go further to formulate linear response as the first term in a functional Taylor expansion to calculate full pairwise joint marginals. However, it may not be obvious to the practitioner how to apply the general formulas of [14]. Our contributions in the present work are (1) the provision of concrete, straightforward formulas for covariance correction that are fast and easy to compute, (2) demonstrations of the success of our method on a wide range of new models, and (3) an accompanying suite of code.

2 Linear response covariance estimation

2.1 Variational Inference

Suppose we observe N data points, denoted by the N-long column vector x, and denote our unobserved model parameters by θ. Here, θ is a column vector residing in some space Θ; it has J subgroups and total dimension D. Our model is specified by a distribution of the observed data given the model parameters—the likelihood $p(x|\theta)$—and a prior distributional belief on the model parameters $p(\theta)$. Bayes' Theorem yields the posterior $p(\theta|x)$.

Mean-field variational Bayes (MFVB) approximates $p(\theta|x)$ by a factorized distribution of the form $q(\theta) = \prod_{j=1}^{J} q(\theta_j)$. q is chosen so that the Kullback-Liebler divergence $\mathrm{KL}(q||p)$ between q and p is minimized. Equivalently, q is chosen so that $E := L + S$, for $L := \mathbb{E}_q[\log p(\theta|x)]$ (the expected log posterior) and $S := -\mathbb{E}_q[\log q(\theta)]$ (the entropy of the variational distribution), is maximized:

$$q^* := \arg\min_q \mathrm{KL}(q||p) = \arg\min_q \mathbb{E}_q\left[\log q(\theta) - \log p(\theta|x)\right] = \arg\max_q E. \tag{1}$$

Up to a constant in θ, the objective E is sometimes called the "evidence lower bound", or the ELBO [2]. In what follows, we further assume that our variational distribution, $q(\theta)$, is in the exponential family with natural parameter η and log partition function A: $\log q(\theta|\eta) = \eta^T\theta - A(\eta)$ (expressed with respect to some base measure in θ). We assume that $p(\theta|x)$ is expressed with respect to the same base measure in θ as for q. Below, we will make only mild regularity assumptions about the true posterior $p(\theta|x)$ and no assumptions about its form.

If we assume additionally that the parameters η^* at the optimum $q^*(\theta) = q(\theta|\eta^*)$ are in the interior of the feasible space, then $q(\theta|\eta)$ may instead be described by the mean parameterization: $m := \mathbb{E}_q\theta$ with $m^* := \mathbb{E}_{q^*}\theta$. Thus, the objective E can be expressed as a function of m, and the first-order condition for the optimality of q^* becomes the fixed point equation

$$\left.\frac{\partial E}{\partial m}\right|_{m=m^*} = 0 \;\Leftrightarrow\; \left.\left(\frac{\partial E}{\partial m} + m\right)\right|_{m=m^*} = m^* \;\Leftrightarrow\; M(m^*) = m^* \text{ for } M(m) := \frac{\partial E}{\partial m} + m. \quad (2)$$

2.2 Linear Response

Let V denote the covariance matrix of θ under the variational distribution $q^*(\theta)$, and let Σ denote the covariance matrix of θ under the true posterior, $p(\theta|x)$:

$$V := \text{Cov}_{q^*}\theta, \qquad \Sigma := \text{Cov}_p\theta.$$

In MFVB, V may be a poor estimator of Σ, even when $m^* \approx \mathbb{E}_p\theta$, i.e., when the marginal estimated means match well [2, 20, 23]. Our goal is to use the MFVB solution and linear response methods to construct an improved estimator for Σ. We will focus on the covariance of the natural sufficient statistic θ, though the covariance of functions of θ can be estimated similarly (see Appendix A).

The essential idea of linear response is to perturb the first-order condition $M(m^*) = m^*$ around its optimum. In particular, define the distribution $p_t(\theta|x)$ as a log-linear perturbation of the posterior:

$$\log p_t(\theta|x) \quad := \quad \log p(\theta|x) + t^T\theta - C(t), \quad (3)$$

where $C(t)$ is a constant in θ. We assume that $p_t(\theta|x)$ is a well-defined distribution for any t in an open ball around 0. Since $C(t)$ normalizes $p_t(\theta|x)$, it is in fact the cumulant-generating function of $p(\theta|x)$, so the derivatives of $C(t)$ evaluated at $t = 0$ give the cumulants of θ. To see why this perturbation may be useful, recall that the second cumulant of a distribution is the covariance matrix, our desired estimand:

$$\Sigma = \text{Cov}_p(\theta) = \left.\frac{d}{dt^T dt}C(t)\right|_{t=0} = \left.\frac{d}{dt^T}\mathbb{E}_{p_t}\theta\right|_{t=0}.$$

The practical success of MFVB relies on the fact that its estimates of the mean are often good in practice. So we assume that $m_t^* \approx \mathbb{E}_{p_t}\theta$, where m_t^* is the mean parameter characterizing q_t^* and q_t^* is the MFVB approximation to p_t. (We examine this assumption further in Section 3.) Taking derivatives with respect to t on both sides of this mean approximation and setting $t = 0$ yields

$$\Sigma = \text{Cov}_p(\theta) \approx \left.\frac{dm_t^*}{dt^T}\right|_{t=0} =: \hat{\Sigma}, \quad (4)$$

where we call $\hat{\Sigma}$ the *linear response variational Bayes* (LRVB) estimate of the posterior covariance of θ.

We next show that there exists a simple formula for $\hat{\Sigma}$. Recalling the form of the KL divergence (see Eq. (1)), we have that $-\mathrm{KL}(q||p_t) = E + t^T m =: E_t$. Then by Eq. (2), we have $m_t^* = M_t(m_t^*)$ for $M_t(m) := M(m) + t$. It follows from the chain rule that

$$\frac{dm_t^*}{dt} = \frac{\partial M_t}{\partial m^T}\bigg|_{m=m_t^*} \frac{dm_t^*}{dt} + \frac{\partial M_t}{\partial t} = \frac{\partial M_t}{\partial m^T}\bigg|_{m=m_t^*} \frac{dm_t^*}{dt} + I, \tag{5}$$

where I is the identity matrix. If we assume that we are at a strict local optimum and so can invert the Hessian of E, then evaluating at $t = 0$ yields

$$\hat{\Sigma} = \frac{dm_t^*}{dt^T}\bigg|_{t=0} = \frac{\partial M}{\partial m}\hat{\Sigma} + I = \left(\frac{\partial^2 E}{\partial m \partial m^T} + I\right)\hat{\Sigma} + I \quad \Rightarrow \quad \hat{\Sigma} = -\left(\frac{\partial^2 E}{\partial m \partial m^T}\right)^{-1}, \tag{6}$$

where we have used the form for M in Eq. (2). So the LRVB estimator $\hat{\Sigma}$ is the negative inverse Hessian of the optimization objective, E, as a function of the mean parameters. It follows from Eq. (6) that $\hat{\Sigma}$ is both symmetric and positive definite when the variational distribution q^* is at least a local maximum of E.

We can further simplify Eq. (6) by using the exponential family form of the variational approximating distribution q. For q in exponential family form as above, the negative entropy $-S$ is dual to the log partition function A [22], so $S = -\eta^T m + A(\eta)$; hence,

$$\frac{dS}{dm} = \frac{\partial S}{\partial \eta^T}\frac{d\eta}{dm} + \frac{\partial S}{\partial m} = \left(\frac{\partial A}{\partial \eta} - m\right)\frac{d\eta}{dm} - \eta(m) = -\eta(m).$$

Recall that for exponential families, $\partial\eta(m)/\partial m = V^{-1}$. So Eq. (6) becomes[1]

$$\hat{\Sigma} = -\left(\frac{\partial^2 L}{\partial m \partial m^T} + \frac{\partial^2 S}{\partial m \partial m^T}\right)^{-1} = -(H - V^{-1})^{-1}, \text{ for } H := \frac{\partial^2 L}{\partial m \partial m^T}. \Rightarrow$$
$$\hat{\Sigma} = (I - VH)^{-1}V. \tag{7}$$

When the true posterior $p(\theta|x)$ is in the exponential family and contains no products of the variational moment parameters, then $H = 0$ and $\hat{\Sigma} = V$. In this case, the mean field assumption is correct, and the LRVB and MFVB covariances coincide at the true posterior covariance. Furthermore, even when the variational assumptions fail, as long as certain mean parameters are estimated exactly, then this formula is also exact for covariances. E.g., notably, MFVB is well-known to provide arbitrarily bad estimates of the covariance of a multivariate normal posterior [2, 11, 20, 23], but since MFVB estimates the means exactly, LRVB estimates the covariance exactly (see Appendix B).

2.3 Scaling the matrix inverse

Eq. (7) requires the inverse of a matrix as large as the parameter dimension of the posterior $p(\theta|x)$, which may be computationally prohibitive. Suppose we are interested in the covariance of parameter sub-vector α, and let z denote the remaining parameters: $\theta = (\alpha, z)^T$. We can partition $\Sigma = (\Sigma_\alpha, \Sigma_{\alpha z}; \Sigma_{z\alpha}, \Sigma_z)$. Similar partitions exist for V and H. If we assume a mean-field factorization $q(\alpha, z) = q(\alpha)q(z)$, then $V_{\alpha z} = 0$. (The variational distributions may factor further as well.) We calculate the Schur complement of $\hat{\Sigma}$ in Eq. (7) with respect to its zth component to find that

$$\hat{\Sigma}_\alpha = \left(I_\alpha - V_\alpha H_\alpha - V_\alpha H_{\alpha z}\left(I_z - V_z H_z\right)^{-1} V_z H_{z\alpha}\right)^{-1} V_\alpha. \tag{8}$$

[1] For a comparison of this formula with the frequentist "supplemented expectation-maximization" procedure see Appendix C.

Here, I_α and I_z refer to α- and z-sized identity matrices, respectively. In cases where $\left(I_z - V_z H_z\right)^{-1}$ can be efficiently calculated (e.g., all the experiments in Section 3; see Fig. (5) in Appendix D), Eq. (8) requires only an α-sized inverse.

3 Experiments

We compare the covariance estimates from LRVB and MFVB in a range of models, including models both with and without conjugacy [2]. We demonstrate the superiority of the LRVB estimate to MFVB in all models before focusing in on Gaussian mixture models for a more detailed scalability analysis.

For each model, we simulate datasets with a range of parameters. In the graphs, each point represents the outcome from a single simulation. The horizontal axis is always the result from an MCMC procedure, which we take as the ground truth. As discussed in Section 2.2, the accuracy of the LRVB covariance for a sufficient statistic depends on the approximation $m_t^* \approx \mathbb{E}_{p_t}\theta$. In the models to follow, we focus on regimes of moderate dependence where this is a reasonable assumption for most of the parameters (see Section 3.2 for an exception). Except where explicitly mentioned, the MFVB means of the parameters of interest coincided well with the MCMC means, so our key assumption in the LRVB derivations of Section 2 appears to hold.

3.1 Normal-Poisson model

Model. First consider a Poisson generalized linear mixed model, exhibiting non-conjugacy. We observe Poisson draws y_n and a design vector x_n, for $n = 1, ..., N$. Implicitly below, we will everywhere condition on the x_n, which we consider to be a fixed design matrix. The generative model is:

$$z_n|\beta,\tau \overset{indep}{\sim} \mathcal{N}\left(z_n|\beta x_n, \tau^{-1}\right), \quad y_n|z_n \overset{indep}{\sim} \text{Poisson}\left(y_n|\exp(z_n)\right), \qquad (9)$$
$$\beta \sim \mathcal{N}(\beta|0,\sigma_\beta^2), \quad \tau \sim \Gamma(\tau|\alpha_\tau,\beta_\tau).$$

For MFVB, we factorize $q\left(\beta,\tau,z\right) = q\left(\beta\right)q\left(\tau\right)\prod_{n=1}^{N} q\left(z_n\right)$. Inspection reveals that the optimal $q\left(\beta\right)$ will be Gaussian, and the optimal $q\left(\tau\right)$ will be gamma (see Appendix D). Since the optimal $q\left(z_n\right)$ does not take a standard exponential family form, we restrict further to Gaussian $q\left(z_n\right)$. There are product terms in L (for example, the term $\mathbb{E}_q\left[\tau\right]\mathbb{E}_q\left[\beta\right]\mathbb{E}_q\left[z_n\right]$), so $H \neq 0$, and the mean field approximation does not hold; we expect LRVB to improve on the MFVB covariance estimate. A detailed description of how to calculate the LRVB estimate can be found in Appendix D.

Results. We simulated 100 datasets, each with 500 data points and a randomly chosen value for μ and τ. We drew the design matrix x from a normal distribution and held it fixed throughout. We set prior hyperparameters $\sigma_\beta^2 = 10$, $\alpha_\tau = 1$, and $\beta_\tau = 1$. To get the "ground truth" covariance matrix, we took 20000 draws from the posterior with the R MCMCglmm package [5], which used a combination of Gibbs and Metropolis Hastings sampling. Our LRVB estimates used the autodifferentiation software JuMP [10].

Results are shown in Fig. (1). Since τ is high in many of the simulations, z and β are correlated, and MFVB underestimates the standard deviation of β and τ. LRVB matches the MCMC standard deviation for all β, and matches for τ in all but the most correlated simulations. When τ gets very high, the MFVB assumption starts to bias the point estimates of τ, and the LRVB standard deviations

[2] All the code is available on our Github repository, rgiordan/LinearResponseVariationalBayesNIPS2015,

start to differ from MCMC. Even in that case, however, the LRVB standard deviations are much more accurate than the MFVB estimates, which underestimate the uncertainty dramatically. The final plot shows that LRVB estimates the covariances of z with β, τ, and $\log \tau$ reasonably well, while MFVB considers them independent.

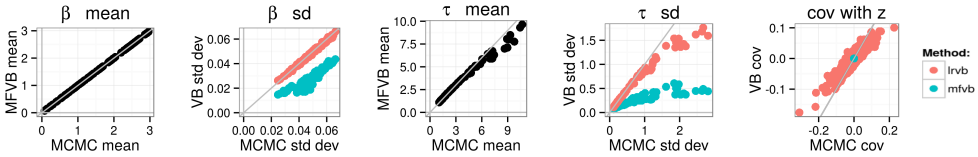


Figure 1: Posterior mean and covariance estimates on normal-Poisson simulation data.

3.2 Linear random effects

Model. Next, we consider a simple random slope linear model, with full details in Appendix E. We observe scalars y_n and r_n and a vector x_n, for $n = 1, ..., N$. Implicitly below, we will everywhere condition on all the x_n and r_n, which we consider to be fixed design matrices. In general, each random effect may appear in multiple observations, and the index $k(n)$ indicates which random effect, z_k, affects which observation, y_n. The full generative model is:

$$y_n|\beta, z, \tau \overset{indep}{\sim} \mathcal{N}\left(y_n|\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right), \quad z_k|\nu \overset{iid}{\sim} \mathcal{N}\left(z_k|0, \nu^{-1}\right),$$
$$\beta \sim \mathcal{N}(\beta|0, \Sigma_\beta), \quad \nu \sim \Gamma(\nu|\alpha_\nu, \beta_\nu), \quad \tau \sim \Gamma(\tau|\alpha_\tau, \beta_\tau).$$

We assume the mean-field factorization $q(\beta, \nu, \tau, z) = q(\beta)\, q(\tau)\, q(\nu) \prod_{k=1}^{K} q(z_n)$. Since this is a conjugate model, the optimal q will be in the exponential family with no additional assumptions.

Results. We simulated 100 datasets of 300 datapoints each and 30 distinct random effects. We set prior hyperparameters to $\alpha_\nu = 2$, $\beta_\nu = 2$, $\alpha_\tau = 2$, $\beta_\tau = 2$, and $\Sigma_\beta = 0.1^{-1} I$. Our x_n was 2-dimensional. As in Section 3.1, we implemented the variational solution using the autodifferentiation software JuMP [10]. The MCMC fit was performed with using MCMCglmm [5].

Intuitively, when the random effect explanatory variables r_n are highly correlated with the fixed effects x_n, then the posteriors for z and β will also be correlated, leading to a violation of the mean field assumption and an underestimated MFVB covariance. In our simulation, we used $r_n = x_{1n} + \mathcal{N}(0, 0.4)$, so that r_n is correlated with x_{1n} but not x_{2n}. The result, as seen in Fig. (2), is that β_1 is underestimated by MFVB, but β_2 is not. The ν parameter, in contrast, is not well-estimated by the MFVB approximation in many of the simulations. Since the LRVB depends on the approximation $m_t^* \approx \mathbb{E}_{p_t}\theta$, its LRVB covariance is not accurate either (Fig. (2)). However, LRVB still improves on the MFVB standard deviation.

3.3 Mixture of normals

Model. Mixture models constitute some of the most popular models for MFVB application [3, 4] and are often used as an example of where MFVB covariance estimates may go awry [2, 20]. Thus, we will consider in detail a Gaussian mixture model (GMM) consisting of a K-component mixture of P-dimensional multivariate normals with unknown component means, covariances, and weights.

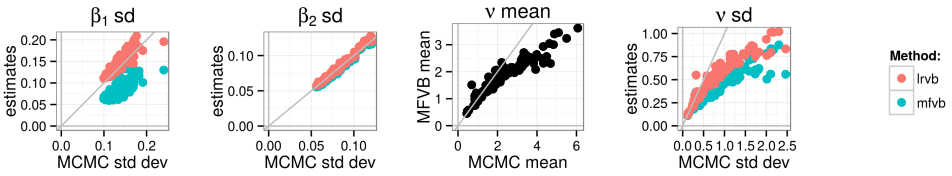


Figure 2: Posterior mean and covariance estimates on linear random effects simulation data.

In what follows, the weight π_k is the probability of the kth component, μ_k is the P-dimensional mean of the kth component, and Λ_k is the $P \times P$ precision matrix of the kth component (so Λ_k^{-1} is the covariance parameter). N is the number of data points, and x_n is the nth observed P-dimensional data point. We employ the standard trick of augmenting the data generating process with the latent indicator variables z_{nk}, for $n = 1, ..., N$ and $k = 1, ..., K$, such that $z_{nk} = 1$ implies $x_n \sim \mathcal{N}(\mu_k, \Lambda_k^{-1})$. So the generative model is:

$$P(z_{nk} = 1) = \pi_k, \quad p(x|\pi, \mu, \Lambda, z) = \prod_{n=1:N} \prod_{k=1:K} \mathcal{N}(x_n|\mu_k, \Lambda_k^{-1})^{z_{nk}} \tag{10}$$

We used diffuse conditionally conjugate priors (see Appendix F for details). We make the variational assumption $q(\mu, \pi, \Lambda, z) = \prod_{k=1}^{K} q(\mu_k) q(\Lambda_k) q(\pi_k) \prod_{n=1}^{N} q(z_n)$. We compare the accuracy and speed of our estimates to Gibbs sampling on the augmented model (Eq. (10)) using the function rnmixGibbs from the R package bayesm. We implemented LRVB in C++, making extensive use of RcppEigen [1]. We evaluate our results both on simulated data and on the MNIST data set [9].

Results. For simulations, we generated $N = 10000$ data points from $K = 2$ multivariate normal components in $P = 2$ dimensions. MFVB is expected to underestimate the marginal variance of μ, Λ, and $\log(\pi)$ when the components overlap since that induces correlation in the posteriors due to the uncertain classification of points between the clusters. We check the covariances estimated with Eq. (7) against a Gibbs sampler, which we treat as the ground truth.[3]

We performed 198 simulations, each of which had at least 500 effective Gibbs samples in each variable—calculated with the R tool effectiveSize from the coda package [16]. The first three plots show the diagonal standard deviations, and the third plot shows the off-diagonal covariances. Note that the off-diagonal covariance plot excludes the MFVB estimates since most of the values are zero. Fig. (3) shows that the raw MFVB covariance estimates are often quite different from the Gibbs sampler results, while the LRVB estimates match the Gibbs sampler closely.

For a real-world example, we fit a $K = 2$ GMM to the $N = 12665$ instances of handwritten 0s and 1s in the MNIST data set. We used PCA to reduce the pixel intensities to $P = 25$ dimensions. Full details are provided in Appendix G. In this MNIST analysis, the Λ standard deviations were under-estimated by MFVB but correctly estimated by LRVB (Fig. (3)); the other parameter standard deviations were estimated correctly by both and are not shown.

[3]The likelihood described in Section 3.3 is symmetric under relabeling. When the component locations and shapes have a real-life interpretation, the researcher is generally interested in the uncertainty of μ, Λ, and π for a particular labeling, not the marginal uncertainty over all possible re-labelings. This poses a problem for standard MCMC methods, and we restrict our simulations to regimes where label switching did not occur in our Gibbs sampler. The MFVB solution conveniently avoids this problem since the mean field assumption prevents it from representing more than one mode of the joint posterior.

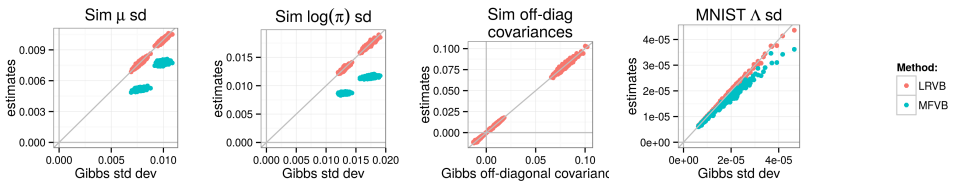


Figure 3: Posterior mean and covariance estimates on GMM simulation and MNIST data.

3.4 Scaling experiments

We here explore the computational scaling of LRVB in more depth for the finite Gaussian mixture model (Section 3.3). In the terms of Section 2.3, α includes the sufficient statistics from μ, π, and Λ, and grows as $O(KP^2)$. The sufficient statistics for the variational posterior of μ contain the P-length vectors μ_k, for each k, and the $(P+1)P/2$ second-order products in the covariance matrix $\mu_k\mu_k^T$. Similarly, for each k, the variational posterior of Λ involves the $(P+1)P/2$ sufficient statistics in the symmetric matrix Λ_k as well as the term $\log|\Lambda_k|$. The sufficient statistics for the posterior of π_k are the K terms $\log\pi_k$.[4] So, minimally, Eq. (7) will require the inverse of a matrix of size $O(KP^2)$. The sufficient statistics for z have dimension $K \times N$. Though the number of parameters thus grows with the number of data points, $H_z = 0$ for the multivariate normal (see Appendix F), so we can apply Eq. (8) to replace the inverse of an $O(KN)$-sized matrix with multiplication by the same matrix. Since a matrix inverse is cubic in the size of the matrix, the worst-case scaling for LRVB is then $O(K^2)$ in K, $O(P^6)$ in P, and $O(N)$ in N.

In our simulations (Fig. (4)) we can see that, in practice, LRVB scales linearly[5] in N and approximately cubically in P across the dimensions considered.[6] The P scaling is presumably better than the theoretical worst case of $O(P^6)$ due to extra efficiency in the numerical linear algebra. Note that the vertical axis of the leftmost plot is on the log scale. At all the values of N, K and P considered here, LRVB was at least as fast as Gibbs sampling and often orders of magnitude faster.

4 Conclusion

The lack of accurate covariance estimates from the widely used mean-field variational Bayes (MFVB) methodology has been a longstanding shortcoming of MFVB. We have demonstrated that in sparse models, our method, linear response variational Bayes (LRVB), can correct MFVB to deliver these covariance estimates in time that scales linearly with the number of data points. Furthermore, we provide an easy-to-use formula for applying LRVB to a wide range of inference problems. Our experiments on a diverse set of models have demonstrated the efficacy of LRVB, and our detailed

[4] Since $\sum_{k=1}^K \pi_k = 1$, using K sufficient statistics involves one redundant parameter. However, this does not violate any of the necessary assumptions for Eq. (7), and it considerably simplifies the calculations. Note that though the perturbation argument of Section 2 requires the parameters of $p(\theta|x)$ to be in the interior of the feasible space, it does not require that the parameters of $p(x|\theta)$ be interior.

[5] The Gibbs sampling time was linearly rescaled to the amount of time necessary to achieve 1000 effective samples in the slowest-mixing component of any parameter. Interestingly, this rescaling leads to increasing efficiency in the Gibbs sampling at low P due to improved mixing, though the benefits cease to accrue at moderate dimensions.

[6] For numeric stability we started the optimization procedures for MFVB at the true values, so the time to compute the optimum in our simulations was very fast and not representative of practice. On real data, the optimization time will depend on the quality of the starting point. Consequently, the times shown for LRVB are only the times to compute the LRVB estimate. The optimization times were on the same order.

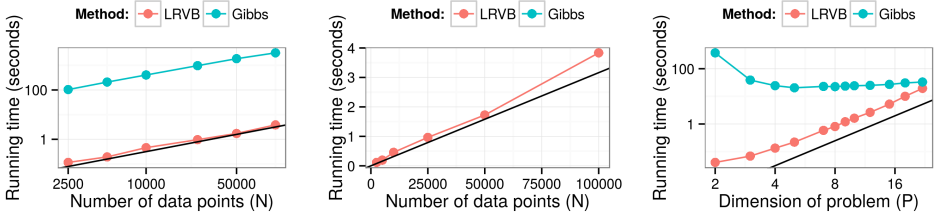


Figure 4: Scaling of LRVB and Gibbs on simulation data in both log and linear scales. Before taking logs, the line in the two lefthand (N) graphs is $y \propto x$, and in the righthand (P) graph, it is $y \propto x^3$.

study of scaling of mixtures of multivariate Gaussians shows that LRVB can be considerably faster than traditional MCMC methods. We hope that in future work our results can be extended to more complex models, including Bayesian nonparametric models, where MFVB has proven its practical success.

Acknowledgments. The authors thank Alex Blocker for helpful comments. R. Giordano and T. Broderick were funded by Berkeley Fellowships.

References

[1] D. Bates and D. Eddelbuettel. Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, 52(5):1–24, 2013.

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

[3] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] J. D. Hadfield. MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R package. *Journal of Statistical Software*, 33(2):1–22, 2010.

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

[7] P. A. d. F. R. Højen-Sørensen, O. Winther, and L. K. Hansen. Mean-field approaches to independent component analysis. *Neural Computation*, 14(4):889–918, 2002.

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

[9] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[11] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003. Chapter 33.

[12] 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.

[13] M. Opper and D. Saad. *Advanced mean field methods: Theory and practice*. MIT press, 2001.

[14] M. Opper and O. Winther. Variational linear response. In *Advances in Neural Information Processing Systems*, 2003.

[15] G. Parisi. *Statistical Field Theory*, volume 4. Addison-Wesley New York, 1988.

[16] M. Plummer, N. Best, K. Cowles, and K. Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, 2006.

[17] H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (statistical methodology)*, 71(2):319–392, 2009.

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

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

[20] 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.

[21] A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, 106(1):25–57, 2006.

[22] 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.

[23] 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.

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

Appendices

You can find this paper, as well as all the code necessary to run the described experiments, in our Github repo, rgiordan/LinearResponseVariationalBayesNIPS2015.

A LRVB estimates of the covariance of functions

In Section 2.2, we derived an estimate of the covariance of the natural sufficient statistics, θ, of our variational approximation, $q(\theta)$. In this section we derive a version of Eq. (7) for the covariance of functions of θ.

We begin by estimating the covariance between θ and a function $\phi(\theta)$. Suppose we have an MFVB solution, $q(\theta)$, to Eq. (1). Define the expectation of $\phi(\theta)$ to be $\mathbb{E}_q\left[\phi(\theta)\right] := f(m)$. This expectation is function of m alone since m completely parameterizes q. As in Eq. (3), we can consider a perturbed log likelihood that also includes $f\left(m\right)$:

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

Using the same reasoning that led to Eq. (4), we will define

$$\Sigma_{\theta\phi} = \mathrm{Cov}_p(\theta, \phi(\theta)) \approx \frac{dm_t^*}{dt_f} =: \hat{\Sigma}_{\theta\phi}$$

We then have the following lemma:

Lemma A.1. *If $\mathbb{E}_q\left[\phi(\theta)\right] =: f(m)$ is a differentiable function of m with gradient ∇f, then*

$$\hat{\Sigma}_{\theta\phi} = \hat{\Sigma}\nabla f$$

Proof. The derivative of the perturbed ELBO, E_t, is given by:

$$\begin{aligned} E_t &:= E + t^T m_f \\ \frac{\partial E_t}{\partial m} &= \frac{\partial E}{\partial m} + \left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right) \end{aligned}$$

The fixed point Eq. (2) then gives:

$$\begin{aligned} M_t\left(m\right) &:= M\left(m\right) + \left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right) \\ \frac{dm_t^*}{dt^T} &= \left.\frac{\partial M_t}{\partial m^T}\right|_{m=m_t^*}\frac{dm_t^*}{dt^T} + \frac{\partial M_t}{\partial t^T} \\ &= \left(\left.\frac{\partial M}{\partial m^T}\right|_{m=m_t^*} + \frac{\partial}{\partial m^T}\left(\begin{array}{cc} I & \nabla f \end{array}\right)\left(\begin{array}{c} t_0 \\ t_f \end{array}\right)\right)\frac{dm^*}{dt^T} + \left(\begin{array}{cc} I & \nabla f \end{array}\right) \end{aligned}$$

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

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

Recalling that

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

We can plug in to see that

$$\hat{\Sigma}_{\theta\phi} = \frac{dm^*}{dt_f} = \hat{\Sigma}\nabla f \tag{11}$$

□

Finally, suppose we are interested in estimating $\text{Cov}_p(\gamma(\theta), \phi(\theta))$, where $g(m) := \mathbb{E}_q[\gamma(\theta)]$. Again using the same reasoning that led to Eq. (4), we will define

$$\Sigma_{\gamma\phi} = \text{Cov}_p(\gamma(\theta), \phi(\theta)) \approx \frac{d\mathbb{E}_q[\gamma(\theta)]}{dt_f} =: \hat{\Sigma}_{\gamma\phi}$$

Proposition A.2. *If $\mathbb{E}_q[\phi(\theta)] = f(m)$ and $\mathbb{E}_q[\gamma(\theta)] = g(m)$ are differentiable functions of m with gradients ∇f and ∇g respectively, then*

$$\hat{\Sigma}_{\gamma\phi} = \nabla g^T \hat{\Sigma} \nabla f$$

Proof. By Lemma A.1 an application of the chain rule,

$$\hat{\Sigma}_{\gamma\phi} = \frac{d\mathbb{E}_q[\gamma(\theta)]}{dt_f} = \frac{dg(m)}{dt_f} = \frac{dg(m)}{dm^T}\frac{dm}{dt_f} = \nabla g^T \hat{\Sigma} \nabla f$$

□

B Exactness of LRVB for multivariate normal means

For any target distribution $p(\theta|x)$, it is well-known that MFVB cannot be used to estimate the covariances between the components of θ. In particular, if q^* is the estimate of $p(\theta|x)$ returned by MFVB, q^* will have a block-diagonal covariance matrix—no matter the form of the covariance of $p(\theta|x)$.

Consider approximating a multivariate Gaussian posterior distribution $p(\theta|x)$ with MFVB. The Gaussian is the unique distribution that is fully determined by its mean and covariance. This posterior arises, for instance, given a multivariate normal likelihood $p(x|\mu) = \prod_{n=1:N}\mathcal{N}(x_n|\mu, S)$ with fixed covariance S and an improper uniform prior on the mean parameter μ. We make the mean

field factorization assumption $q(\mu) = \prod_{d=1:D} q(\mu_d)$, where D is the total dimension of μ. This fact is often used to illustrate the shortcomings of MFVB [2, 20, 23]. In this case, it is well known that the MFVB posterior means are correct, but the marginal variances are underestimated if S is not diagonal. However, since the posterior means are correctly estimated, the LRVB approximation in Eq. (7) is in fact an equality. That is, for this model, $\hat{\Sigma} = dm_t/dt^T = \Sigma$ exactly.

In order to prove this result, we will rely on the following lemma.

Lemma B.1. *Consider a target posterior distribution characterized by $p(\theta|x) = \mathcal{N}(\theta|\mu, \Sigma)$, where μ and Σ may depend on x, and Σ is invertible. Let $\theta = (\theta_1, \ldots, \theta_J)$, and consider a MFVB approximation to $p(\theta|x)$ that factorizes as $q(\theta) = \prod_j q(\theta_j)$. Then the variational posterior means are the true posterior means; i.e. $m_j = \mu_j$ for all j between 1 and J.*

Proof. The derivation of MFVB for the multivariate normal can be found in Section 10.1.2 of [2]; we highlight some key results here. Let $\Lambda = \Sigma^{-1}$. Let the j index on a row or column correspond to θ_j, and let the $-j$ index correspond to $\{\theta_i : i \in [J] \setminus j\}$. E.g., for $j = 1$,

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{1,-1} \\ \Lambda_{-1,1} & \Lambda_{-1,-1} \end{bmatrix}.$$

By the assumption that $p(\theta|x) = \mathcal{N}(\theta|\mu, \Sigma)$, we have

$$\begin{aligned} &\log p(\theta_j|\theta_{i\in[J]\setminus j}, x) \\ &= -\frac{1}{2}(\theta_j - \mu_j)^T \Lambda_{jj}(\theta_j - \mu_j) + (\theta_j - \mu_j)^T \Lambda_{j,-j}(\theta_{-j} - \mu_{-j}) + C, \end{aligned} \tag{12}$$

where the final term is constant with respect to θ_j. It follows that

$$\begin{aligned} \log q_j^*(\theta_j) &= \mathbb{E}_{q_i^*: i\in[J]\setminus j} \log p(\theta, x) + C \\ &= -\frac{1}{2}\theta_j^T \Lambda_{jj}\theta_j + \theta_j \mu_j \Lambda_{jj} - \theta_j \Lambda_{j,-j}(\mathbb{E}_{q^*}\theta_{-j} - \mu_{-j}). \end{aligned}$$

So

$$q_j^*(\theta_j) = \mathcal{N}(\theta_j|m_j, \Lambda_{jj}^{-1}),$$

with mean parameters

$$m_j = \mathbb{E}_{q_j^*}\theta_j = \mu_j - \Lambda_{jj}^{-1}\Lambda_{j,-j}(m_{-j} - \mu_{-j}) \tag{13}$$

as well as an equation for $\mathbb{E}_{q^*}\theta^T\theta$.

Note that Λ_{jj} must be invertible, for if it were not, Σ would not be invertible.

The solution $m = \mu$ is a unique stable point for Eq. (13), since the fixed point equations for each

j can be stacked and rearranged to give

$$
\begin{aligned}
m - \mu &= -\begin{bmatrix} 0 & \Lambda_{11}^{-1}\Lambda_{12} & \cdots & \Lambda_{11}^{-1}\Lambda_{1(J-1)} & \Lambda_{11}^{-1}\Lambda_{1J} \\ \vdots & & \ddots & & \vdots \\ \Lambda_{JJ}^{-1}\Lambda_{J1} & \Lambda_{JJ}^{-1}\Lambda_{J2} & \cdots & \Lambda_{JJ}^{-1}\Lambda_{J(J-1)} & 0 \end{bmatrix} (m - \mu) \\
&= -\begin{bmatrix} \Lambda_{11}^{-1} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \Lambda_{JJ}^{-1} \end{bmatrix} \begin{bmatrix} 0 & \Lambda_{12} & \cdots & \Lambda_{1(J-1)} & \Lambda_{1J} \\ \vdots & & \ddots & & \vdots \\ \Lambda_{J1} & \Lambda_{J2} & \cdots & \Lambda_{J(J-1)} & 0 \end{bmatrix} (m - \mu) \Leftrightarrow \\
0 &= \begin{bmatrix} \Lambda_{11} & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \ddots & & 0 \\ \vdots & & & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & \Lambda_{JJ} \end{bmatrix} (m - \mu) + \\
&\quad \begin{bmatrix} 0 & \Lambda_{12} & \cdots & \Lambda_{1(J-1)} & \Lambda_{1J} \\ \vdots & & \ddots & & \vdots \\ \Lambda_{J1} & \Lambda_{J2} & \cdots & \Lambda_{J(J-1)} & 0 \end{bmatrix} (m - \mu) \Leftrightarrow \\
0 &= \Lambda (m - \mu) \Leftrightarrow \\
m &= \mu.
\end{aligned}
$$

The last step follows from the assumption that Σ (and hence Λ) is invertible. It follows that μ is the unique stable point of Eq. (13).

□

Proposition B.2. *Assume we are in the setting of Lemma B.1, where additionally μ and Σ are on the interior of the feasible parameter space. Then the LRVB covariance estimate exactly captures the true covariance, $\hat{\Sigma} = \Sigma$.*

Proof. Consider the perturbation for LRVB defined in Eq. (3). By perturbing the log likelihood, we change both the true means μ_t and the variational solutions, m_t. The result is a valid density function since the original μ and Σ are on the interior of the parameter space. By Lemma B.1, the MFVB solutions are exactly the true means, so $m_{t,j} = \mu_{t,j}$, and the derivatives are the same as well. This means that the first term in Eq. (7) is not approximate, i.e.

$$\frac{dm_t}{dt^T} = \frac{d}{dt^T}\mathbb{E}_{p_t}\theta = \Sigma_t,$$

It follows from the arguments above that the LRVB covariance matrix is exact, and $\hat{\Sigma} = \Sigma$.

□

C Comparison with supplemented expectation-maximization

The result in Appendix B about the multivariate normal distribution draws a connection between LRVB corrections and the "supplemented expectation-maximization" (SEM) method of [12]. SEM

is an asymptotically exact covariance correction for the EM algorithm that transforms the full-data Fisher information matrix into the observed-data Fisher information matrix using a correction that is formally similar to Eq. (7). In this section, we argue that this similarity is not a coincidence; in fact the SEM correction is an asymptotic version of LRVB with two variational blocks, one for the missing data and one for the unknown parameters.

Although LRVB as described here requires a prior (unlike SEM, which supplements the MLE), the two covariance corrections coincide when the full information likelihood is approximately log quadratic and proportional to the posterior, $p(\theta|x)$. This might be expected to occur when we have a large number of independent data points informing each parameter—i.e., when a central limit theorem applies and the priors do not affect the posterior. In the full information likelihood, some terms may be viewed as missing data, whereas in the Bayesian model the same terms may be viewed as latent parameters, but this does not prevent us from formally comparing the two methods.

We can draw a term-by-term analogy with the equations in [12]. We denote variables from the SEM paper with a superscript "SEM" to avoid confusion. MFVB does not differentiate between missing data and parameters to be estimated, so our θ corresponds to $(\theta^{SEM}, Y_{mis}^{SEM})$ in [12]. SEM is an asymptotic theory, so we may assume that $(\theta^{SEM}, Y_{mis}^{SEM})$ have a multivariate normal distribution, and that we are interested in the mean and covariance of θ^{SEM}.

In the E-step of [12], we replace Y_{mis}^{SEM} with its conditional expectation given the data and other θ^{SEM}. This corresponds precisely to Eq. (13), taking $\theta_j = Y_{mis}^{SEM}$. In the M-step, we find the maximum of the log likelihood with respect to θ^{SEM}, keeping Y_{mis}^{SEM} fixed at its expectation. Since the mode of a multivariate normal distribution is also its mean, this, too, corresponds to Eq. (13), now taking $\theta_j = \theta^{SEM}$.

It follows that the MFVB and EM fixed point equations are the same; i.e., our M is the same as their M^{SEM}, and our $\partial M/\partial m$ of Eq. (5) corresponds to the transpose of their DM^{SEM}, defined in Eq. (2.2.1) of [12]. Since the "complete information" corresponds to the variance of θ^{SEM} with fixed values for Y_{OBS}^{SEM}, this is the same as our $\Sigma_{q^*,11}$, the variational covariance, whose inverse is I_{oc}^{-1}. Taken all together, this means that equation (2.4.6) of [12] can be re-written as our Eq. (7).

$$\begin{aligned} V^{SEM} =& I_{oc}^{-1}\left(I - DM^{SEM}\right)^{-1} \Rightarrow \\ \Sigma =& V\left(I - \left(\frac{\partial M}{\partial m^T}\right)^T\right)^{-1} = \left(I - \frac{\partial M}{\partial m^T}\right)^{-1} V \end{aligned}$$

D Normal-Poisson details

In this section, we use this model to provide a detailed, step-by-step description of a simple LRVB analysis.

The full joint distribution for the model in Eq. (9) is

$$\begin{aligned} \log p\left(y, z, \beta, \tau\right) = & \sum_{n=1}^{N}\left(-\frac{1}{2}\tau z_n^2 + x_n\tau\beta z_n - \frac{1}{2}x_n^2\tau\beta^2 - \frac{1}{2}\log\tau\right) \\ & + \sum_{n=1}^{N}\left(-\exp\left(z_n\right) + z_n y_n\right) - \frac{1}{2\sigma_\beta^2}\beta^2 + \left(\alpha_\tau - 1\right)\log\tau - \beta_\tau\tau + C \end{aligned}$$

We find a mean-field approximation under the factorization $q\left(\beta, \tau, z\right) = q\left(\beta\right)q\left(\tau\right)\prod_{n=1}^{N} q\left(z_n\right)$. By inspection, the log joint is quadratic in β, so the optimal $q\left(\beta\right)$ will be Gaussian [2]. Similarly,

the log joint is a function of τ only via τ and $\log \tau$, so the optimal $q(\tau)$ will be gamma. However, the joint does not take a standard exponential family form in z_n:

$$\log p(z_n|y, \beta, \tau) = (x_n \tau \beta + y_n) z_n - \frac{1}{2} \tau z_n^2 - \exp(z_n) + C$$

The difficulty is with the term $\exp(z_n)$. So we make the further restriction that

$$q(z_n) = \mathcal{N}(\cdot) = q\left(z_n; \mathbb{E}[z_n], \mathbb{E}\left[z_n^2\right]\right).$$

Fortunately, the troublesome term has an analytic expectation, as a function of the mean parameters, under this variational posterior:

$$\mathbb{E}_q[\exp(z_n)] = \exp\left(\mathbb{E}_q[z_n] + \frac{1}{2}\left(\mathbb{E}_q\left[z_n^2\right] - \mathbb{E}_q[z_n]^2\right)\right).$$

We can now write the variational distribution in terms of the following mean parameters:

$$m = \left(\mathbb{E}_q[\beta], \mathbb{E}_q\left[\beta^2\right], \mathbb{E}_q[\tau], \mathbb{E}_q[\log \tau], \mathbb{E}_q[z_1], \mathbb{E}_q\left[z_1^2\right], ..., \mathbb{E}_q[z_N], \mathbb{E}_q\left[z_N^2\right]\right)^T.$$

Calculating the LRVB covariance consists of roughly four steps:

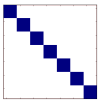
1. finding the MFVB optimum q^*,
2. computing the covariance V of q^*,
3. computing H, the Hessian of $L(m)$, for q^*, and
4. computing the matrix inverse and solving $(I - VH)^{-1} V$.

For step (1), the LRVB correction is agnostic as to how the optimum is found. In our experiments below, we follow a standard coordinate ascent procedure for MFVB [2]. We analytically update $q(\beta)$ and $q(\tau)$. Given $q(\beta)$ and $q(\tau)$, finding the optimal $q(z)$ becomes N separate two-dimensional optimization problems; there is one dimension for each of the mean parameters $\mathbb{E}_q[z_n]$ and $\mathbb{E}_q\left[z_n^2\right]$. In our examples, we solved these problems sequentially using IPOPT [21].

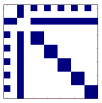
To compute V for step (2), we note that by the mean-field assumption, β, τ, and z_n are independent, so V is block diagonal. Since we have chosen convenient variational distributions, the mean parameters have known covariance matrices. For example, from standard properties of the normal distribution, $\text{Cov}\left(\beta, \beta^2\right) = 2\mathbb{E}_q[\beta]\left(\mathbb{E}_q\left[\beta^2\right] - \mathbb{E}_q[\beta]^2\right)$.

For step (3), the mean parameters for β and τ co-occur with each other and with all the z_n, so these four rows of H are expected to be dense. However, the mean parameters for z_n never occur with each other, so the bulk of H—the $2N \times 2N$ block corresponding to the mean parameters of z—will be block diagonal (Fig. (5b)). The Hessian of $L(m)$ can be calculated analytically, but we used the autodifferentiation software JuMP [10].

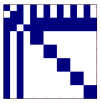
Finally, for step (4), we use the technique in Section 2.3 to exploit the sparsity of V and H (Fig. (5c)) in calculating $(I - VH)^{-1}$.



(a) MFVB covariance V



(b) Hessian matrix H



(c) $(I - VH)$

Figure 5: Sparsity patterns for $\hat{\Sigma} = (I - VH)^{-1}$ using the model in Eq. (9), $n = 5$ (white = 0)

E Random effects model details

As introduced in Section 3.2, our model is:

$$\begin{aligned} y_n|\beta, z, \tau &\overset{indep}{\sim} \mathcal{N}\left(\beta^T x_n + r_n z_{k(n)}, \tau^{-1}\right) \\ z_k|\nu &\overset{iid}{\sim} \mathcal{N}\left(0, \nu^{-1}\right) \end{aligned}$$

With the priors:

$$\begin{aligned} \beta &\sim \mathcal{N}\left(0, \Sigma_\beta\right) \\ \nu &\sim \Gamma\left(\alpha_\nu, \beta_\nu\right) \\ \tau &\sim \Gamma\left(\alpha_\tau, \beta_\tau\right) \end{aligned}$$

We will make the following mean field assumption:

$$q\left(\beta, z, \tau, \nu\right) = q\left(\nu\right) q\left(\tau\right) q\left(\beta\right) \prod_{k=1}^{K} q\left(z_k\right)$$

We have $n \in \{1, ..., N\}$, and $k \in \{1, ..., K\}$, and $k\left(n\right)$ matches an observation n to a random effect k, allowing repeated observations of a random effect. The full joint log likelihood is:

$$\begin{aligned} \log p\left(y_n|z_{k(n)}, \tau, \beta\right) &= -\frac{\tau}{2}\left(y_n - \beta^T x_n - r_n z_{k(n)}\right)^2 + \frac{1}{2}\log\tau + C \\ \log p\left(z_k|\nu\right) &= -\frac{\nu}{2}z_k^2 + \frac{1}{2}\log\nu + C \\ \log p\left(\beta\right) & -\frac{1}{2}\text{trace}\left(\Sigma_\beta^{-1}\beta\beta^T\right) + C \\ \log p\left(\tau\right) &= \left(\alpha_\tau - 1\right)\log\tau - \beta_\tau\tau + C \\ \log p\left(\nu\right) &= \left(\alpha_\nu - 1\right)\log\nu - \beta_\nu\nu + C \\ \log p\left(y, \tau, \beta, z\right) &= \sum_{n=1}^{N}\log p\left(y_n|z_{k(n)}, \tau, \beta\right) + \sum_{k=1}^{K}\log p\left(z_k|\nu\right) + \\ & \log p\left(\beta\right) + \log p\left(\nu\right) + \log p\left(\tau\right) \end{aligned}$$

Expanding the first term of the conditional likelihood of y_n gives

$$\begin{aligned} & -\frac{\tau}{2}\left(y_n - \beta^T x_n - r_n z_{k(n)}\right)^2 \\ &= -\frac{\tau}{2}\left(y_n^2 - 2y_n x_n^T\beta - 2y_n r_n z_{n(k)} + \text{trace}\left(x_n x_n^T\beta\beta^T\right) + r_n^2 z_{k(n)}^2 + 2r_n x_n^T\beta z_{k(n)}\right) \end{aligned}$$

By grouping terms, we can see that the mean parameters will be

$$\begin{aligned} q(\beta) &= q\left(\beta; \mathbb{E}_q[\beta], \mathbb{E}_q\left[\beta\beta^T\right]\right) \\ q(z_k) &= q\left(z_k; \mathbb{E}_q[z_k], \mathbb{E}_q\left[z_k^2\right]\right) \\ q(\tau) &= q(\tau; \mathbb{E}_q[\tau], \mathbb{E}_q[\log\tau]) \\ q(\nu) &= q(\nu; \mathbb{E}_q[\nu], \mathbb{E}_q[\log\nu]) \end{aligned}$$

It follows that the optimal variational distributions are $q(\beta)$ =multivariate normal, $q(z_k)$ =univariate normal, and $q(\tau)$ and $q(\nu)$ will be gamma. We performed standard coordinate ascent on these distributions [2].

As in Section 3.1, we implemented this model in the autodifferentiation software JuMP [10]. This means conjugate coordinate updates were easy, since the natural parameters corresponding to a mean parameters are the first derivatives of the log likelihood with respect to the mean parameters. For example, denoting the log likelihood at step s by L_s, the update for $q_{s+1}(z_k)$ will be:

$$\log q_{s+1}(z_k) = \frac{\partial \mathbb{E}_q[L_s]}{\partial \mathbb{E}_q[z_k]} z_k + \frac{\partial \mathbb{E}_q[L_s]}{\partial \mathbb{E}_q[z_k^2]} z_k^2 + C$$

Given the partial derivatives of L_s with respect to the mean parameters, the updated mean parameters for z_k can be read off directly using standard properties of the normal distribution.

The variational covariance matrices are all standard. We can see that H will have nonzero terms in general (for example, the three-way interaction $\mathbb{E}_q[\tau]\,\mathbb{E}_q\left[z_{k(n)}\right]\mathbb{E}_q[\beta]$), and that LRVB will be different from MFVB. As usual in our models, H is sparse, and we can easily apply the technique in section Section 2.3 to get the covariance matrix excluding the random effects, z.

F Multivariate normal mixture details

In this section we derive the basic formulas needed to calculate Eq. (7) for a finite mixture of normals, which is the model used in Section 3. We will follow the notation introduced in Section 3.3.

Let each observation, x_n, be a $P \times 1$ vector. We will denote the Pth component of the nth observation x_n, with a similar pattern for z and μ. We will denote the p, qth entry in the matrix Λ_k as $\Lambda_{k,pq}$. The data generating process is as follows:

$$\begin{aligned} P(x|\mu,\pi,\Lambda) &= \prod_{n=1}^{N} P(x_n|z_n,\mu,\Lambda) \prod_{k=1}^{K} P(z_{nk}|\pi_k) \\ \log P(x_n|z_n,\mu,\Lambda) &= \sum_{n=1}^{N} z_{nk} \log\phi_k(x_n) + C \\ \log\phi_k(x) &= -\frac{1}{2}(x-\mu_k)^T \Lambda_k (x-\mu_k) + \frac{1}{2}\log|\Lambda_k| + C \\ \log P(z_{nk}|\pi_k) &= \sum_{k=1}^{K} z_{nk}\log\pi_k + C \end{aligned}$$

It follows that the log posterior is given by

$$\begin{aligned}\log P(z,\mu,\pi,\Lambda|x) &= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk}\left(\log\pi_k - \frac{1}{2}\left(x_n-\mu_k\right)^T\Lambda_k\left(x_n-\mu_k\right) + \frac{1}{2}\log|\Lambda_k|\right) + \\ &\quad \sum_{k=1}^{K}\log p(\mu_k) + \sum_{k=1}^{K}\log p(\Lambda_k) + \log p(\pi) + C\end{aligned}$$

We used a multivariate normal prior for μ_k, a Wishart prior for Λ_k, and a Dirichlet prior for π. In the simulations described in Section 3.3, we used the following prior parameters for the VB model:

$$\begin{aligned}p(\mu_k) &= \mathcal{N}\left(0_P, \text{diag}_P(0.01)^{-1}\right) \\ p(\Lambda_k) &= \text{Wishart}(\text{diag}_P(0.01), 1) \\ p(\pi) &= \text{Dirichlet}(5_K)\end{aligned}$$

Here, $\text{diag}_P(a)$ is a P-dimensional diagonal matrix with a on the diagonal, and 0_P is a length P vector of the value 0, with a similar definition for 5_K. Unfortunately, the function we used for the MCMC calculations, rnmixGibbs in the package bayesm, uses a different form for the μ_k prior. Specifically, rnmixGibbs uses the prior

$$p_{MCMC}\left(\mu_k|\,\Lambda_k\right) = \mathcal{N}(0, a^{-1}\Lambda_k^{-1})$$

where a is a scalar. There is no way to exactly match $p_{MCMC}(\mu_k)$ to $p(\mu_k)$, so we simply set $a = 0.01$. Since our datasets are all reasonably large, the prior was dominated by the likelihood, and we found the results extremely insensitive to the prior on μ_k, so this discrepancy is of no practical importance.

The parameters μ_k, Λ_k, π, and z_n will each be given their own variational distribution. For q_{μ_k} we will use a multivariate normal distribution; for q_{Λ_k} we will us a Wishart distirbution; for q_π we will use a Dirichlet distribution; for q_{z_n} we will use a Multinoulli (a single multinomial draw). These are all the optimal variational choices given the mean field assumption and the conditional conjugacy in the model.

The sufficient statistics for μ_k are all terms of the form μ_{kp} and $\mu_{kp}\mu_{kq}$. Consequently, the sub-vector of θ corresponding to μ_k is

$$\theta_{\mu_k} = \begin{pmatrix} \mu_{k1} \\ \vdots \\ \mu_{kp} \\ \mu_{k1}\mu_{k1} \\ \mu_{k1}\mu_{k2} \\ \vdots \\ \mu_{kP}\mu_{kP} \end{pmatrix}$$

We will only save one copy of $\mu_{kp}\mu_{kq}$ and $\mu_{kq}\mu_{kp}$, so θ_{μ_k} has length $P + \frac{1}{2}\left(P+1\right)P$. For all the parameters, we denote the complete stacked vector without a k subscript:

$$\theta_\mu = \begin{pmatrix} \theta_{\mu_1} \\ \vdots \\ \theta_{\mu_K} \end{pmatrix}$$

The sufficient statistics for Λ_k are all the terms $\Lambda_{k,pq}$ and the term $\log|\Lambda_k|$. Again, since Λ is symmetric, we do not keep redundant terms, so θ_{Λ_k} has length $1+\frac{1}{2}(P+1)P$. The sufficient statistic for π is the K-vector $(\log\pi_1,...,\log\pi_K)$. The sufficient statistics for z are simply the $N\times K$ values z_{nk} themselves.

In terms of Section 2.3, we have

$$\begin{aligned}\alpha &= \begin{pmatrix}\theta_\mu\\ \theta_\Lambda\\ \theta_\pi\end{pmatrix}\\ z &= \begin{pmatrix}\theta_z\end{pmatrix}\end{aligned}$$

That is, we are primarily interested in the covariance of the sufficient statistics of μ, Λ, and π. The latent variables z are nuisance parameters.

To put the log likelihood in terms useful for LRVB, we must express it in terms of the sufficient statistics, taking into account the fact the θ vector does not store redundant terms (e.g. it will only keep Λ_{ab} for $a<b$ since Λ is symmetric).

$$\begin{aligned}&-\frac{1}{2}(x_n-\mu_k)^T\Lambda_k(x_n-\mu_k)\\ &= -\frac{1}{2}\text{trace}\left(\Lambda_k(x_n-\mu_k)(x_n-\mu_k)^T\right)\\ &= -\frac{1}{2}\sum_a\sum_b\left(\Lambda_{k,ab}(x_{n,a}-\mu_{k,a})(x_{n,b}-\mu_{k,b})\right)\\ &= -\frac{1}{2}\sum_a\sum_b\left(\Lambda_{k,ab}\mu_{k,a}\mu_{k,b}-\Lambda_{k,ab}x_{n,a}\mu_{k,b}-\Lambda_{k,ab}x_{n,b}\mu_{k,a}+\Lambda_{k,ab}x_{n,a}x_{n,b}\right)\\ &= -\frac{1}{2}\sum_a\Lambda_{k,aa}\left(\mu_k^2\right)^a+\sum_a\Lambda_{k,aa}x_{n,a}\mu_{k,a}-\frac{1}{2}\sum_a\Lambda_{k,aa}\left(x_n^2\right)^2-\\ &\quad\frac{1}{2}\sum_{a\neq b}\Lambda_{k,ab}\mu_{k,a}\mu_{k,b}+\sum_{a\neq b}\Lambda_{k,ab}x_{n,a}\mu_{k,b}-\frac{1}{2}\sum_{a\neq b}\Lambda_{k,ab}x_{n,a}x_{n,b}\\ &= -\frac{1}{2}\sum_a\Lambda_{k,aa}\left(\mu_k^2\right)^a+\sum_a\Lambda_{k,aa}x_{n,a}\mu_{k,a}-\frac{1}{2}\sum_a\Lambda_{k,aa}\left(x_n^2\right)^2-\\ &\quad\sum_{a<b}\Lambda_{k,ab}\mu_{k,a}\mu_{k,b}+\sum_{a<b}\Lambda_{k,ab}\left(x_{n,a}\mu_{k,b}+x_{n,b}\mu_{k,a}\right)-\sum_{a<b}\Lambda_{k,ab}x_{n,a}x_{n,b}\end{aligned}$$

The MFVB updates and covariances in V are all given by properties of standard distributions. To compute the LRVB corrections, it only remains to calculate the Hessian, H. These terms can be read directly off the posterior. First we calculate derivatives with respect to components of μ.

$$\begin{aligned}\frac{\partial^2 H}{\partial\mu_{k,a}\partial\Lambda_{k,ab}} &= \sum_i z_{nk}x_{n,b}\\ \frac{\partial^2 H}{\partial(\mu_{k,a}\mu_{k,b})\partial\Lambda_{k,ab}} &= -\left(\frac{1}{2}\right)^{1(a=b)}\sum_n z_{nk}\\ \frac{\partial^2 H}{\partial\mu_{k,a}\partial z_{nk}} &= \sum_b\Lambda_{k,ab}x_{n,b}\\ \frac{\partial^2 H}{\partial(\mu_{k,a}\mu_{k,b})\partial z_{nk}} &= -\left(\frac{1}{2}\right)^{1(a=b)}\Lambda_{k,ab}\end{aligned}$$

All other μ derivatives are zero. For Λ,

$$\begin{aligned}
\frac{\partial^2 H}{\partial \Lambda_{k,ab} \partial z_{nk}} &= -\left(\frac{1}{2}\right)^{1(a=b)} (x_{n,a} x_{n,b} - \mu_{k,a} x_{n,b} - \mu_{k,b} x_{n,a} + \mu_{k,a} \mu_{k,b}) \\
\frac{\partial^2 H}{\partial \log |\Lambda_k| \, \partial z_{nk}} &= \frac{1}{2}
\end{aligned}$$

The remaining Λ derivatives are zero. The only nonzero second derivatives for $\log \pi$ are to Z and are given by

$$\frac{\partial^2 H}{\partial \log \pi_j \partial z_{nk}} = 1$$

Note in particular that $H_{zz} = 0$, allowing efficient calculation of Eq. (8).

G MNIST details

For a real-world example, we applied LRVB to the unsupervised classification of two digits from the MNIST dataset of handwritten digits. We first preprocess the MNIST dataset by performing principle component analysis on the training data's centered pixel intensities and keeping the top 25 components. For evaluation, the test data is projected onto the same 25-dimensional subspace found using the training data.

We then treat the problem of separating handwritten 0s from 1s as an unsupervised clustering problem. We limit the dataset to instances labeled as 0 or 1, resulting in 12665 training and 2115 test points. We fit the training data as a mixture of multivariate Gaussians. Here, $K = 2$, $P = 25$, and $N = 12665$. Then, keeping the μ, Λ, and π parameters fixed, we calculate the expectations of the latent variables z in Eq. (10) for the test set. We assign test set data point x_n to whichever component has maximum a posteriori expectation. We count successful classifications as test set points that match their cluster's majority label and errors as test set points that are different from their cluster's majority label. By this measure, our test set error rate was 0.08. We stress that we intend only to demonstrate the feasibility of LRVB on a large, real-world dataset rather than to propose practical methods for modeling MNIST.