

# Beyond prediction: A framework for inference with variational approximations in mixture models

T. Westling\*

Department of Statistics  
University of Washington

T. H. McCormick

Departments of Statistics & Sociology  
University of Washington

November 22, 2017

## Abstract

Variational inference is a popular method for estimating model parameters and conditional distributions in hierarchical and mixed models, which arise frequently in many settings in the health, social, and biological sciences. Variational inference in a frequentist context works by approximating intractable conditional distributions with a tractable family and optimizing the resulting lower bound on the log-likelihood. The variational objective function is typically less computationally intensive to optimize than the true likelihood, enabling scientists to fit rich models even with extremely large datasets. Despite widespread use, little is known about the general theoretical properties of estimators arising from variational approximations to the log-likelihood, which hinders their use in inferential statistics. In this paper we connect such estimators to profile M-estimation, which enables us to provide regularity conditions for consistency and asymptotic normality of variational estimators. Our theory also motivates three methodological improvements to variational inference: estimation of the asymptotic model-robust covariance matrix, a one-step correction that improves estimator efficiency, and an empirical assessment of consistency. We evaluate the proposed results using simulation studies and data on marijuana use from the National Longitudinal Study of Youth.

*Keywords:* generalized linear mixed models; profile M-estimation.

---

\*The authors gratefully acknowledge grant 62389-CS-YIP from the United States Army Research Office, grants SES-1559778 and DMS-1737673 from the National Science Foundation, and grant number K01 HD078452 from the National Institute of Child Health and Human Development (NICHD).

# 1 Introduction

Thanks to rapid improvements in data availability and user-friendly tools for data manipulation and storage, researchers from an ever broader set of scientific disciplines now routinely analyze extremely complex, high-dimensional data. However, computing parameter estimates using stalwart statistical techniques such as maximum likelihood and Markov chain Monte Carlo can be a challenge in these settings and is often a bottleneck in practice. In these situations, researchers turn to computationally efficient approximations.

*Variational approximations* are one method of approximating a likelihood function or posterior distribution that are increasingly popular across a range of scientific fields. In public health, for example, Lee & Wand (2016) used a variational approximation to estimate a model for overall and hospital-specific trends in cesarean section rates. In statistical genetics, Raj et al. (2014) used a variational approximation to a multinomial model of allele frequencies across populations of individuals. O’Connor et al. (2010) used a variational approximation to a model of demographics and lexical choice in geo-tagged Twitter data.

Despite their popularity, variational approximations do not typically come with guarantees about the statistical properties of the resulting estimator. This drawback is particularly problematic when a scientist would like to interpret a parameter estimate, in which case estimator consistency is crucial, or report a confidence interval, in which case good coverage rates rely on the ability to accurately estimate the sampling distribution of the estimator. In this paper, we address the problem of inference using variational approximations. We show that, in a wide range of parametric mixture models, well-established theory from profile  $M$ -estimation provides an asymptotic lens through which we may understand the properties of parameter estimates resulting from variational approximations to the log-likelihood. Using the  $M$ -estimation framework, we derive conditions for consistency and asymptotic normality of variational estimators.

We also propose three methodological improvements to variational estimators. First, we provide a consistent estimator of the asymptotic covariance matrix of variational estimators. Second, we introduce a one-step correction to the variational estimator that improves large-sample statistical efficiency. Third, we develop an empirical evaluation of estimator consistency when the theoretical calculations are intractable. We demonstrate the impor-

tance of these methodological advances with two logistic mixture models of marijuana use by age among participants in the National Longitudinal Survey of Youth (NLSY).

The remainder of the paper is organized as follows. This section formally defines the class of models and variational estimators we study. Section 2 connects variational estimation to profile  $M$ -estimation and states our theoretical results. Section 3 presents our three methodological contributions. Section 4 evaluates our method using simulated data and demonstrates an application to the NLSY. Section 5 presents a discussion.

Code to replicate all of the empirical analyses in this paper are available at [https://github.com/tedwestling/variational\\_asymptotics](https://github.com/tedwestling/variational_asymptotics).

## 1.1 Variational estimators

In this paper, we consider inference for a Euclidean parameter  $\theta$  in a parametric mixture model  $p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x, z) d\mu(z)$ , where the marginal likelihood  $p_\theta(x)$  is computationally expensive to compute. Parametric mixture models have been used in a variety of scientific contexts. For example, mixed-membership models are a type of mixture model that have been used to model text (Blei et al. 2003), social networks (Airoldi et al. 2008), population genetics (Pritchard et al. 2000), and scientific collaborations (Erosheva et al. 2004).

Mixture models have been used in conjunction with both Bayesian and frequentist inferential frameworks. In a frequentist setting, maximum likelihood (ML) estimation comes with guarantees of asymptotic efficiency and methods of conducting inference for many models. These guarantees provide a degree of assurance for scientists that the point estimates and uncertainty intervals will behave in predictable ways.

ML estimation can, however, be computationally burdensome. When the integral in  $p_\theta(x)$  must be approximated numerically, the cost of this computation increases exponentially with the dimension of the domain  $\mathcal{Z}$  of the latent variable since  $p_\theta(x, z)$  needs to be evaluated at sufficiently many points to accurately approximate the integral. This computational burden is a significant barrier for researchers who want to develop tailored mixture models to flexibly represent the dependencies in their data. As a result, a variety of approximate methods have been developed as alternatives to maximum likelihood.

Variational inference is an approximate method based on optimizing a lower bound for

the original objective function. The variational lower bound is designed to reduce the dimension of the required numerical integration, thereby improving computational efficiency. Variational inference can be used in a frequentist context to approximate the log-likelihood and in a Bayesian context to approximate the full posterior distribution. In this paper we focus on the former. We will refer to estimators of  $\theta$  resulting from optimizing a variational approximation to the log-likelihood as *variational estimators*.

Before providing formal definitions, we distinguish between two key aspects of the variational approximation. First, we can evaluate the properties of the optimizer of the variational lower bound. Second, we could consider the tightness of the variational lower bound to the true objective function. These questions are related. Demonstrating tightness of the lower bound is one way to control the difference between the true and variational optimizers, for example. However, a tight variational lower bound is not a necessary condition for good behavior of the variational estimator, and indeed does not hold in many settings where the variational estimator performs well. In this paper, we address the former of these two components, i.e. the properties of the optimizer of the variational lower bound.

We now move to a formal definition of variational estimation. More thorough explanations can be found in, for example, Wainwright & Jordan (2008) or Ormerod & Wand (2010). Let  $X_1, \dots, X_n$  be observed  $p$ -variate data generated independently and identically from a distribution  $P_0$  on a sample space  $\mathcal{X}$ . Let  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  be a statistical model where  $\Theta$  is an open subset of  $\mathbb{R}^d$  and each  $P_\theta$  has a density  $p_\theta(x) = \int_{\mathcal{Z}} p_\theta(x, z) d\mu(z)$ . Here  $\mu$  is a dominating measure on  $\mathcal{Z} \subseteq \mathbb{R}^k$ , the sample space of the latent random variable  $Z$ .

We are most interested in cases where  $p_\theta(x)$  cannot be written in closed-form in terms of elementary functions, as in many generalized linear mixed models (McCulloch & Neuhaus 2001) and non-linear hierarchical models (Davidian & Giltinan 1995, Goldstein 2011). In these cases, calculating the log-likelihood of the observed data,  $\sum_{i=1}^n \log p_\theta(X_i)$ , and its derivatives with respect to  $\theta$  requires numerical integration. When the dimension of the latent variable is large, these numerical integrals are computationally expensive.

In order to motivate variational approximations, it is useful to represent the true parameter  $\theta_0$  and the true conditional distribution of the latent variable  $\pi_{\theta_0}(z \mid x)$  as joint

maximizers of a Kullback-Liebler divergence:

$$(\theta_0, \pi_{\theta_0}) \equiv \arg \max_{\theta \in \Theta, q \in \mathcal{Q}_0} E_{P_0} \left[ E_q \left[ \log \frac{p_\theta(X, Z)}{q(Z | X)} \mid X \right] \right], \quad (1)$$

where  $\mathcal{Q}_0$  is the class of all densities dominated by  $\mu$ . The expectation-maximization (EM) algorithm can be motivated by (1) by replacing the unknown  $P_0$  with the empirical distribution and alternating between optimization over  $\theta$  and  $q$ . However, if the marginal likelihood  $p_\theta(x)$  cannot be written in terms of elementary functions, then neither can  $\pi_\theta$ , and hence the EM algorithm still requires numerical integration.

To construct a variational approximation to the log-likelihood we replace the optimization over  $\mathcal{Q}_0$  in (1) with an optimization over a smaller *variational family* of distributions  $\mathcal{Q}$ . For example,  $\mathcal{Q}$  could consist of all independent products over each dimension of  $z$  (known as mean-field variational inference), all multivariate Gaussian distributions, or all independent Gaussian distributions. For simplicity we will assume throughout that  $\mathcal{Q}$  is indexed by a finite-dimensional Euclidean parameter  $\psi \in \Psi$ , so that every  $Q \in \mathcal{Q}$  can be identified with a density  $q(\cdot; \psi)$ . We note that in some cases even when  $\mathcal{Q}$  is a semi-parametric family, it can be shown that the optimal  $q$  lies in a parametric sub-family with a known form, so that our results can still be applied (see, e.g. Section 5.3 of Wainwright & Jordan (2008)). For families where this doesn't apply, our theory could be extended to incorporate semiparametric  $\mathcal{Q}$ .

Given  $\mathcal{Q}$  and  $\Psi$ , the variational estimator of  $\theta$ , which we will denote  $\hat{\theta}_n$ , and the variational conditional estimators  $\hat{\psi}_n$  are the joint maximizers of the following objective function:

$$(\hat{\theta}_n, \hat{\psi}_n) \equiv \arg \max_{\theta \in \Theta, \psi_n \in \Psi^n} \sum_{i=1}^n E_{\psi_i} \left[ \log \frac{p(X_i, Z | \theta)}{q(Z; \psi_i)} \right] = \arg \max_{\theta \in \Theta, \psi_n \in \Psi^n} \mathcal{L}_n(\theta, \psi_n; \mathbf{X}_n). \quad (2)$$

A crucial piece of motivation for our work is that, since  $\mathcal{L}_n$  is typically not proportional to the log-likelihood, it is not clear what the asymptotic properties are of the variational estimator  $\hat{\theta}_n$ . In many circumstances, the variational estimator is used for prediction. In such cases, scientists can evaluate the quality of the variational approximation using cross-validation or another held-out data technique. If a scientist, however, would like to interpret the point estimator (or, critically, its uncertainty) produced by a variational approximation, not knowing the properties of the estimator is a substantial hindrance. In

particular, we would like to know whether  $\hat{\theta}_n$  is consistent and, if it is consistent, what the asymptotic distribution of  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  is.

There are limited results regarding the general theoretical properties of variational estimators. Specific cases have been studied in depth, yielding positive results regarding the consistency of variational estimators for Gaussian mixture models (Wang & Titterton 2006), exponential family models with missing values (Wang & Titterton 2004), Poisson mixed models as the cluster size and number of clusters both diverge (Hall, Ormerod & Wand 2011, Hall, Pham, Wand & Wang 2011), Markovian models with missing values (Hall et al. 2002), and stochastic block models for social networks (Bickel et al. 2013).

Of particular note are the foundational papers by Hall, Ormerod & Wand (2011) and Hall, Pham, Wand & Wang (2011), which derive sharp asymptotics for Poisson regression with random cluster intercepts as both the number of clusters and observations per cluster diverge. Our work is distinct from these results in two ways. First, we provide results at a general level, and in particular permitting an arbitrary dimension of the latent variables, rather than for a specific model. Second, we focus on the asymptotic regime where the number of clusters is diverging, but the number of observations per cluster is stochastically bounded.

## 2 Variational approximations and M-Estimation

In this section, we demonstrate the connection between M-estimators and variational inference. The key for this connection is using a profile version of the variational objective function. Viewing variational inference in this way unlocks a deep and broad set of theoretical results developed for M-estimators. We make this connection explicit in this section and then, in Section 3, demonstrate how these theoretical results can be used to develop new methods for scientific practice.

### 2.1 Variational estimation as M-Estimation

We will study the general properties of the variational estimator  $\hat{\theta}_n$  through the lens of  $M$ -estimation. An  $M$ -estimator of a parameter  $\theta$  is the maximizer of a data-dependent

objective function

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^n m(\theta; X_i).$$

From (2) we can see that

$$\mathcal{L}_n(\theta, \boldsymbol{\psi}_n; \mathbf{X}_n) = \sum_{i=1}^n v(\theta, \psi_i; X_i) \text{ for } v(\theta, \psi; x) = E_\psi[\log \frac{p_\theta(x, z)}{q(z; \psi)}].$$

Applying the theory of  $M$ -estimation to the vector  $(\theta, \boldsymbol{\psi}_n)$  with  $m(\cdot) = v(\cdot)$  is complicated due to the dependence on  $\psi_i$ , which are known as *incidental* parameters specific to each data point. The  $\theta$ , in contrast, are *structural* parameters shared across all data (Lancaster 2000). Hall, Ormerod & Wand (2011) dealt with this problem for Poisson mixed models by assuming the cluster size was growing with the number of observations, so that the incidental parameters effectively became structural. In our more general setting we could analogously assume that each observed data  $X_i$  is composed of replicates  $X_{i1}, \dots, X_{im}$  and let  $m$  grow with  $n$ . However, this would limit the applicability of our results to only cases where clusters are very large. Since, in practice, clusters are often small, we instead apply  $M$ -estimation to the *profiled* variational objective.

In order to use the  $M$ -estimation framework for the variational estimator  $\hat{\theta}_n$ , we will express the optimization defined in (2) as a two-stage procedure, where first  $\mathcal{L}_n$  is optimized with respect to  $\boldsymbol{\psi}_n$  for each fixed  $\theta$  and second this *profiled* function is optimized with respect to  $\theta$ . Specifically, we define the profiled single-data objective function

$$m(\theta; x) \equiv \sup_{\psi \in \Psi} v(\theta, \psi; x), \text{ so that } \hat{\theta}_n = \arg \max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^n m(\theta; X_i).$$

This form now falls within the  $M$ -estimator framework. We can, therefore, use the existing, well-studied, asymptotic theory of  $M$ -estimators to better understand the asymptotic properties of variational estimators. In the subsequent sections, we show that, using this representation, the theory for  $M$ -estimators yields general results for consistency and asymptotic normality for variational estimators.

## 2.2 Consistency

We first explore consistency using the  $M$ -estimator representation of the variational estimator. An important point that we will return to later is that, depending on the model and

approximation, the estimator based on the variational lower bound may not be consistent for the truth. Hence in what follows, we refer to  $\bar{\theta}$  as the limit of  $\hat{\theta}_n$ , so that  $\bar{\theta} = \theta_0$  if and only if  $\hat{\theta}_n$  is consistent.

The true objective function  $M_0(\theta) = E_{P_0}[m(\theta; X)]$  governs the asymptotic properties of the variational estimator  $\hat{\theta}_n$ . Under regularity conditions,  $\hat{\theta}_n \rightarrow_{P_0} \arg \max_{\theta} M_0(\theta)$ , so that if  $M_0$  is uniquely maximized at  $\theta_0$  then  $\hat{\theta}_n$  is consistent for  $\theta_0$ . Let  $\hat{\psi}(\theta; x) = \arg \max_{\psi \in \Psi} v(\theta, \psi; x)$ , which we will assume exists and is unique for all  $\theta$  and  $x$ . Thus we can write  $m(\theta; x) = v(\theta, \hat{\psi}(\theta; x); x)$  and we have the following consistency result.

**Theorem 1.** *Suppose the function  $M_0(\theta) = E_{P_0}[v(\theta, \hat{\psi}(\theta; X); X)]$  attains a finite global maximum at  $\bar{\theta}$  and conditions (A1)-(A3) hold. Then  $\hat{\theta}_n \rightarrow_{P_0} \bar{\theta}$ .*

Regularity conditions (A1)-(A3) justify the application of Theorem 5.14 of van der Vaart (2000) and are provided in the supplementary material. In practice, it is often not possible to derive  $M_0(\theta)$  in closed form, which prevents a theoretical assessment of consistency of the variational estimator. This is the situation, for instance, in many generalized mixed models. In Section 3.3, we propose a test for consistency that does not require explicit derivation of  $M_0(\theta)$ .

## 2.3 Asymptotic normality

If the variational estimator  $\hat{\theta}_n$  is consistent for  $\bar{\theta}$  and additional regularity conditions hold then  $\sqrt{n}(\hat{\theta}_n - \bar{\theta}) \rightarrow_{d, P_0} N(0, V(\bar{\theta}))$  where  $V(\theta)$  is the sandwich covariance.

**Theorem 2.** *Suppose  $\hat{\theta}_n \rightarrow_{P_0} \bar{\theta}$  and conditions (B1)-(B5) hold. Then  $\sqrt{n}(\hat{\theta}_n - \bar{\theta}) \rightarrow_d N_d(0, V(\bar{\theta}))$  where  $V(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-1}$  for*

$$A(\theta) = E_{P_0} [D_{\theta}^2 m(\theta; X)] \quad (3)$$

$$B(\theta) = E_{P_0} [(D_{\theta} m(\theta; X))(D_{\theta} m(\theta; X))^T]. \quad (4)$$

Conditions (B1)-(B5) guarantee the profiled objective function  $m(\theta; X)$  satisfies the requirements of van der Vaart (2000) Theorem 5.23. In the next section we provide formulas for estimating the matrices  $A$  and  $B$  regardless of whether  $m(\theta; X)$  is known explicitly.



### 3 Practical tools for inference with variational estimators

We now propose three methodological innovations based on the asymptotic results from the previous section. First, we demonstrate how to leverage asymptotic normality to enhance uncertainty estimators. Second, we show that a one-step correction can be applied to improve the efficiency of the variational estimator. Finally, we address the consistency issue mentioned in the previous section, providing a way to test the consistency of a variational estimator when theoretical calculations based on the above formulation are intractable.

#### 3.1 Sandwich covariance estimation

We now discuss computation of consistent covariance estimators. Recall that in practice,  $m(\theta; X)$  is often not available in closed form. Fortunately, the derivatives of  $m(\theta; X)$  can be expressed in terms of the derivatives of  $v(\theta, \psi; X)$ , which are always available. In particular,  $v(\theta, \psi; X)$  is a result of the model and variational family used, meaning the chain rule can be applied so that the asymptotic variance can be estimated whether or not  $m(\theta; X)$  is tractable. Concerning  $D_\theta m(\theta; x)$ , which appears in equation (4), since  $\hat{\psi}(\theta; x)$  maximizes  $v$  for fixed  $\theta, x$ ,  $D_\theta m(\theta; x) = D_\theta v(\theta, \psi; x)|_{\psi=\hat{\psi}(\theta; x)}$ . For  $D_\theta^2 m(\theta; x)$  in equation (3),

$$D_\theta^2 m(\theta; x) = \left[ D_{\theta\theta}^2 v - D_{\theta\psi}^2 v (D_{\psi\psi}^2 v)^{-1} D_{\theta\psi}^2 v^T \right]_{\psi=\hat{\psi}(\theta; x)},$$

where we abbreviate  $v(\theta, \psi; X)$  as  $v$  for presentation. Replacing the appropriate derivatives in the definition of  $V(\theta)$  with the above expressions and the population expectations with empirical ones gives a way to calculate the asymptotic covariance only knowing  $v(\theta, \psi; X)$  and its derivatives (which can be calculated numerically), as opposed to  $m(\theta; x)$ , the computation of which involves optimization.

Thus,  $\hat{A}_n(\hat{\theta}_n)^{-1} \hat{B}_n(\hat{\theta}_n) \hat{A}_n(\hat{\theta}_n)^{-1} \rightarrow_{P_0} V(\bar{\theta})$  where

$$\hat{A}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left[ D_{\theta\theta}^2 v - D_{\theta\psi}^2 v (D_{\psi\psi}^2 v)^{-1} D_{\theta\psi}^2 v^T \right]_{\psi=\hat{\psi}_i, x=X_i}, \quad (5)$$

$$\hat{B}_n(\theta) = \frac{1}{n} \sum_{i=1}^n \left[ (D_\theta v) (D_\theta v)^T \right]_{\psi=\hat{\psi}_i, x=X_i}. \quad (6)$$

Equations (5) and (6) provide a formula for constructing an asymptotic covariance matrix for the variational estimator  $\hat{\theta}_n$ . This covariance can be used to construct asymptotically calibrated Wald intervals, regions, and hypothesis tests about  $\theta_0$  if  $\bar{\theta} = \theta_0$ . Furthermore, the sandwich covariance is model-robust in the sense that it is valid even if  $P_0 \notin \mathcal{P}$ .

For an MLE under correct model specification,  $A(\theta) = B(\theta)$  and the asymptotic covariance reduces to  $A(\theta)^{-1}$ , the inverse Fisher information matrix. In this case the sandwich covariance is only needed for model-robust uncertainty estimation. However, when  $m$  is not proportional to the log-likelihood, as is often true with variational inference,  $A$  and  $B$  are not necessarily equal even under correct model specification. Therefore the sandwich covariance is necessary even if  $P_0 \in \mathcal{P}$ .

### 3.2 One-step correction

The variational estimator  $\hat{\theta}_n$  is not guaranteed to be asymptotically efficient since the variational objective function need not be proportional to the log-likelihood. Hence while Wald-type intervals, regions, and tests using the sandwich estimator proposed in the last section will be asymptotically valid, they may be suboptimal since  $\hat{\theta}_n$  may have larger asymptotic variance than the MLE. In these cases a one-step correction to the variational estimator yields a more efficient estimator.

The one-step estimator is  $\hat{\theta}_n^{(1)} = \hat{\theta}_n - I_n(\hat{\theta}_n)^{-1} S_n(\hat{\theta}_n)$ , where  $l(\theta; x) = \log p_\theta(x)$  and

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^n D_\theta l(\theta; X_i), \quad I_n(\theta) = \frac{1}{n} \sum_{i=1}^n (D_\theta l(\theta; X_i))(D_\theta l(\theta; X_i))^T$$

are the score and observed information at  $\theta$ . Under regularity conditions  $\sqrt{n}(\hat{\theta}_n^{(1)} - \theta_0) \rightarrow_{d, \theta_0} N(0, I(\theta_0)^{-1})$  for  $I(\theta_0)$  the Fisher information matrix, which is the same asymptotic distribution as the maximum likelihood estimator or posterior mean.

Computing  $S_n$  and  $I_n$  require numerical integration in the same way that computing the MLE would. Indeed, the one-step correction is a single step of a Newton-Raphson algorithm for finding the MLE starting at  $\hat{\theta}_n$ . Thus, unlike finding the exact MLE, this one-step procedure only requires a single calculation of these quantities, so requires less computation than finding the exact MLE. However, in some cases the one-step correction may not be computationally feasible for the same reasons that computing the MLE is not.

### 3.3 An empirical test of the consistency of variational estimators

In many cases, including generalized linear mixed models, neither  $\hat{\psi}(\theta; x)$ ,  $m(\theta; x)$ , nor  $M_0(\theta)$  are available analytically. This presents a challenge not present in the classical  $M$ -estimation scenario and seriously undermines the goal of theoretically evaluating the consistency of variational estimators. Simulation studies could be used to assess consistency for any particular fixed, known truth, but would be computationally burdensome.

Here we propose a method for evaluating the consistency of a variational estimator at a single fixed parameter value  $\theta^*$  when  $m(\theta; x)$  is intractable. The method is based on numerically evaluating whether  $E_{\theta^*}[D_{\theta}m(\theta; X)]$  is equal to 0 at  $\theta = \theta^*$ , which is the primary condition for consistency proscribed by Theorem 1. Our method is as follows.

1. Fix  $\theta^*$  and  $b$  very large (for instance  $10^4$  or  $10^5$ ).
2. For  $j = 1, \dots, b$ :
  - (a) Simulate  $X_j^* \sim P_{\theta^*}$ .
  - (b) Find  $\psi_j^* = \hat{\psi}(\theta^*; X_j^*)$  by numerically optimizing  $\psi \mapsto v(\theta^*, \psi; X_j^*)$ .
  - (c) Evaluate  $G_j^* = D_{\theta}v|_{\theta^*, \psi_j^*, X_j^*}$ .
3. Test the null hypothesis that  $E_{\theta^*}[G_j^*] = 0$  either using independent  $t$ -tests on each component or Hotelling's  $T^2$  test on the entire vector.

If the test rejects the null hypothesis then the variational estimator cannot be consistent; if not then one can be arbitrarily certain (with large enough  $b$ ) that the mean score is zero at  $\theta^*$ . If a weakly significant  $p$ -value is found and it is unclear what to conclude, the experiment could be repeated with a larger  $b$ .

This method is a necessary, but not sufficient test of consistency. As we explain more below, asymptotically we expect our method to have few false negatives (indication that the estimator is inconsistent when it is actually consistent) but possibly false positives (indications that the estimator is consistent when it is actually inconsistent). The first reason for potential false positives is that  $E_{\theta^*}[G_j^*] = 0$  is a necessary but not sufficient condition for consistency. Even if its gradient is zero,  $\theta^*$  it need not be a global maxima of

the objective function. The second reason for potential false positives is that the method can only assess consistency at a single parameter value  $\theta^*$  rather than on the entirety of the parameter space. Typically one will first use the variational algorithm to estimate  $\hat{\theta}_n$ , then use this method to assess consistency at  $\theta^* = \hat{\theta}_n$ . If the estimator is consistent for every  $\theta$  in a neighborhood of  $\theta_0$  then for  $n$  large enough  $\hat{\theta}_n$  will be in that neighborhood and the method will not indicate inconsistency. On the other hand if  $\hat{\theta}_n \rightarrow_P \bar{\theta} \neq \theta_0$  then this method is approximately assessing whether the algorithm is consistent near  $\bar{\theta}$ . If the variational algorithm is consistent at  $\bar{\theta}$  but not at  $\theta_0$  then the method would indicate that the estimator is consistent when in fact it is not. Despite the possibility of false positives, we do not know of any other practical ways to assess consistency of variational estimators when the limit objective is not available in closed form.

## 4 Empirical evaluations of variational estimators

In this section we empirically evaluate our methods in three parametric latent variable models. The results indicate that variational estimators are not always consistent: in the first example the estimator is consistent for the entire vector, in the second it is consistent for some parameters and not for others, and in the third it is not consistent for any parameters. The second example also demonstrates that even when the variational estimator is consistent, it is not necessarily efficient. In either case the sandwich covariance matrix provides good confidence interval coverage rates and the one-step correction improves efficiency. The empirical evaluation of consistency correctly identifies inconsistency of the parameter vector as a whole, but not always inconsistency of individual parameters.

### 4.1 Example 1: consistent and efficient estimation

For our first case, we consider data simulated from an exponential mixture model. Suppose that we observe pairs  $(X_{i,1}, X_{i,2})$  generated independently and identically with  $X_{i,1} \sim \text{Exp}(\lambda_1)$ , and  $X_{i,2}|Z_i \sim \text{Exp}(\lambda_1 Z_i)$  where  $Z_i$  is a latent variable with  $Z_i \sim \text{Exp}(\lambda_2)$ . The parameter vector is  $\theta = (\log \lambda_1, \log \lambda_2) \in \mathbb{R}^2$ .

For the purpose of demonstrating our theoretical method of assessing consistency it is

illustrative to consider a variational class of conditional distributions that does not include the true conditional. Suppose the variational class is taken to be all log-normal distributions parametrized by  $\psi = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Psi$ . In this case it turns out the optimal variational parameters can be found in closed form (we provide this derivation in the supplementary material). Thus the profile objective function  $m(\theta; x) = v(\theta, \hat{\psi}(\theta; x); x)$  can be written explicitly, and doing so it turns out that  $m(\theta; x) = \log p_\theta(x) + c$ : the profiled variational objective is equal up to a constant to the log likelihood. Therefore, despite the fact that the variational class *does not contain* the true conditional distribution, the variational estimator  $\hat{\theta}_n$  is equivalent to the MLE.

Next we conduct a simulation study to empirically evaluate the intervals and regions created using the sandwich covariance. Since the variational estimator is equivalent to the MLE in this case the inverse Fisher information gives correct coverage under correct model specification (though not under incorrect specification). For this simulation study we use the mean-field variational estimator obtained by taking the variational class to be all independent products over  $Z_1, \dots, Z_n$ , which includes the true conditional. We construct intervals and regions using the variational estimator and our sandwich covariance matrix. For a point of comparison we also compute the variational Bayes posterior distribution using conjugate prior distributions and the mean-field class of variational posteriors.

Table 1 shows the empirical coverage of marginal and joint 95% regions using variational Bayes and the sandwich covariance. We simulated 1000 data sets of size  $n = 1000$  both from the model described above ( $P \in \mathcal{P}$ ) and with the model misspecified such that  $X_{i1} \sim \text{Gamma}(3, 3\lambda_1)$  ( $P \notin \mathcal{P}$ ). The variational posterior generally results in undercoverage, with coverage under the correct model specification thirteen and thirty percentage points below the nominal 95% level for both individual parameters and 29 points below for the joint coverage. Using these intervals would almost certainly bias substantive scientific conclusions. The proposed sandwich covariance, however, has correct coverage under both correct and incorrect model specification.

Table 1: Coverage of 95% marginal and joint regions in the exponential mixture model.

Method	Parameter	Coverage ( $P_0 \in \mathcal{P}$ )	Coverage ( $P_0 \notin \mathcal{P}$ )
Variational Bayes	$\lambda_1$	0.82	0.99
	$\lambda_2$	0.65	0.73
	$(\lambda_1, \lambda_2)$	0.66	0.78
Sandwich covariance	$\lambda_1$	0.95	0.94
	$\lambda_2$	0.95	0.95
	$(\lambda_1, \lambda_2)$	0.95	0.95

## 4.2 Example 2: consistent but inefficient estimation

In our second and third empirical evaluations of our proposed methodology, we consider mixed effects logistic regression models; first with only random intercepts, and second with random intercepts, slopes, and quadratic terms. In both cases we use data on marijuana use in adolescents in the United States from the National Longitudinal Survey of Youth 1997 (Bureau of Labor Statistics, U.S. Department of Labor 2013). The data consists of  $n = 8660$  youth with at most yearly interviews from 1997 to 2012, with the number of interviews per youth ranging from four to sixteen. For youth  $i$ 's  $j$ th interview we consider the binary outcome  $Y_{ij}$  of whether the youth used marijuana in the thirty days preceding the interview. We focus on understanding the relationship between marijuana use, age, and sex. Since our goal is to understand the properties of variational estimators, we use the data, along with “known” parameter values, to simulate outcomes. This way we can assess the accuracy of parameter estimates and coverage of uncertainty intervals. We also analyze the data with observed outcomes. We mention this analysis briefly in this section and provide additional results in the supplementary materials.

First we consider logistic regression with random intercepts. Let  $Z_i$  be a random intercept controlling each youth's overall propensity for marijuana use,  $SEX_i$  be an indicator that the youth is male, and  $AGE_{ij}$  be youth  $i$ 's age at interview  $j$ . Denote

$p_{ij} = P(Y_{ij} = 1|Z_i, SEX_i, AGE_{ij})$ . Then our first model for marijuana usage is

$$\text{logit}(p_{ij}) = \begin{cases} Z_i + \beta_0 + \beta_1(AGE_{ij}/35) + \beta_2(AGE_{ij}/35)^2, & SEX_i = 0 \\ Z_i + \beta_3 + \beta_4(AGE_{ij}/35) + \beta_5(AGE_{ij}/35)^2, & SEX_i = 1. \end{cases}$$

Each youth's outcomes  $Y_{i1}, \dots, Y_{in_i}$  are assumed conditionally independent given  $Z_i$ , and we model  $Z_i$  as IID  $N(0, \sigma^2)$ . The parameter vector is  $\theta = (\beta, \log(\sigma^2))$ . The inclusion of the quadratic effect of age is important because we expect that marijuana usage peaks some time in young adulthood and decreases thereafter. This model form is similar to that used in the analysis of age-crime curves (Fabio et al. 2011).

To estimate  $\theta$  we consider a variational class of conditional distributions over  $Z_n$  consisting of all independent Gaussian distributions. This is known as a Gaussian variational approximation (GVA). The variational parameters are  $\psi_i = (m_i, \log s_i)$  for  $i = 1, \dots, n$ ,  $m_i$  being the mean and  $s_i$  the standard deviation of the variational conditional distribution of  $\gamma_i$ . The variational objective involves one-dimensional numerical integrals. To optimize the variational objective function we use a variational EM algorithm. In this case it is not possible to express the profiled objective function explicitly.

To evaluate our methods, we conduct a simulation study based on the NLSY data. For each of 1000 simulations, we draw a bootstrap sample of youth. Conditional on these youth's age and sex data we simulate  $Y_1, \dots, Y_n$  from the model, treating the variational estimate  $\hat{\theta}_n$  for the data as the "true" parameter value  $\theta_0$ . We then estimate the model parameters using maximum likelihood with the R package `lme4` (Bates et al. 2015), the Gaussian variational approximation, and the one-step correction to the variational estimator. We also estimate the asymptotic covariance matrix for each method. Finally, we use our proposed method to assess consistency of the variational algorithm at the estimated parameter value.

We first examine the accuracy of point estimates for regression fixed effects. All three estimators concentrate on the true values of the fixed effects  $\beta_0$  through  $\beta_5$  (box plots are provided in the supplementary material). Table 2 shows the variance of the estimators for each of the seven model parameters. The variational estimator has slightly larger variance than the MLE, but the one-step correction nearly matches the variance of the MLE. Thus, as we asserted theoretically, the one step correction is efficient as long as the variational

Table 2: Estimator variances in the logistic regression with random intercepts simulation.

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\log(\sigma^2)$
MLE	0.16	1.68	1.05	0.12	1.25	0.77	$8.6 \times 10^{-3}$
GVA	0.23	1.90	1.19	0.19	1.46	0.91	0.61
One-step correction to GVA	0.17	1.67	1.04	0.13	1.26	0.78	0.60

estimator is consistent, even when the variational estimator is inefficient.

Moving now to the random intercept variance, the MLE concentrates on the true variance component,  $\log(\sigma^2)$ , while the variational estimator and one-step correction do not. Our empirical assessment of consistency described in Section 3.3 correctly identifies this inconsistency. The multivariate Hotelling test rejected in every simulation with  $p < 10^{-16}$ , correctly indicating that the population mean gradient of the entire parameter vector was significantly different from zero. Additionally, no more than 2.5% of the marginal  $t$ -tests rejected at the 0.01 level for each of the fixed effect, in line with their apparent consistency, while every one of the 1000 simulations rejected the marginal  $t$ -test for the variance parameter (the maximal  $p$ -value for the variance parameter was less than  $10^{-16}$ ). These results are better than what is guaranteed theoretically, since theory does not guarantee that the marginal  $t$ -tests will accurately reflect the consistency or inconsistency of individual parameters.

We now move to examining uncertainty intervals. Table 3 shows the estimated coverage of marginal 95% Wald-type confidence intervals of the model parameters for each the three estimators. The coverage of the confidence intervals of the linear and quadratic age fixed effects using the sandwich covariance for the variational estimator and the inverse Fisher information for the one-step correction are within the Monte Carlo error of the nominal 95%. The variational confidence intervals for the sex-specific intercepts  $\beta_0$  and  $\beta_3$  are too small at 90%, likely because of the underestimation of  $\sigma^2$  of the random intercept. The coverage of the confidence intervals of  $\log \sigma^2$  is close to zero for the variational estimator and one-step correction, which is not surprising given that the estimator is inconsistent. The coverage of  $\log \sigma^2$  is not shown for the MLE because `lme4` does not provide an interval



Table 3: Coverage of 95% confidence intervals in the logistic regression with random intercepts simulation.

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$	$\log(\sigma^2)$
Maximum likelihood	0.94	0.94	0.94	0.95	0.95	0.95	–
GVA + sandwich covariance	0.90	0.94	0.94	0.90	0.94	0.94	0.09
One-step correction to GVA	0.93	0.94	0.94	0.91	0.95	0.95	0.02

for this parameter.

We also estimated the model parameters using our one-step correction for the observed outcomes from the NLSY. Figure 1 shows the estimated mean curves as a function of age for both models and both females and males. The average male and female from the random quadratic model have slightly faster increases, peak at younger ages, and decrease earlier than the average male and female from the random intercept model. In both models the average male has higher overall probability and slightly later peak usage: in the random intercept model, the estimated peak female usage probability occurs at 21.3 years (95% CI: [18.1, 24.5]), peak male usage at 22.2 years (95% CI: [19.9, 24.6]). in the random quadratic model, the estimated peak female usage probability occurs at 17.9 years (95% CI: [15.9, 20.0]), peak male usage at 19.1 years (95% CI: [17.3, 20.8]).

### 4.3 Example 3: inconsistent estimation

We now alter the model presented in the last section to include random slopes and quadratic terms for each youth. The random intercept model may not accurately capture the dependence structure of a single subject’s marijuana use over time. The random intercepts model implies an exchangeable marginal correlation structure, which is unrealistic given the longitudinal nature of the data. A more realistic model allows random slopes and quadratic terms as well, so that the latent variable  $Z_i$  now has three components. For the

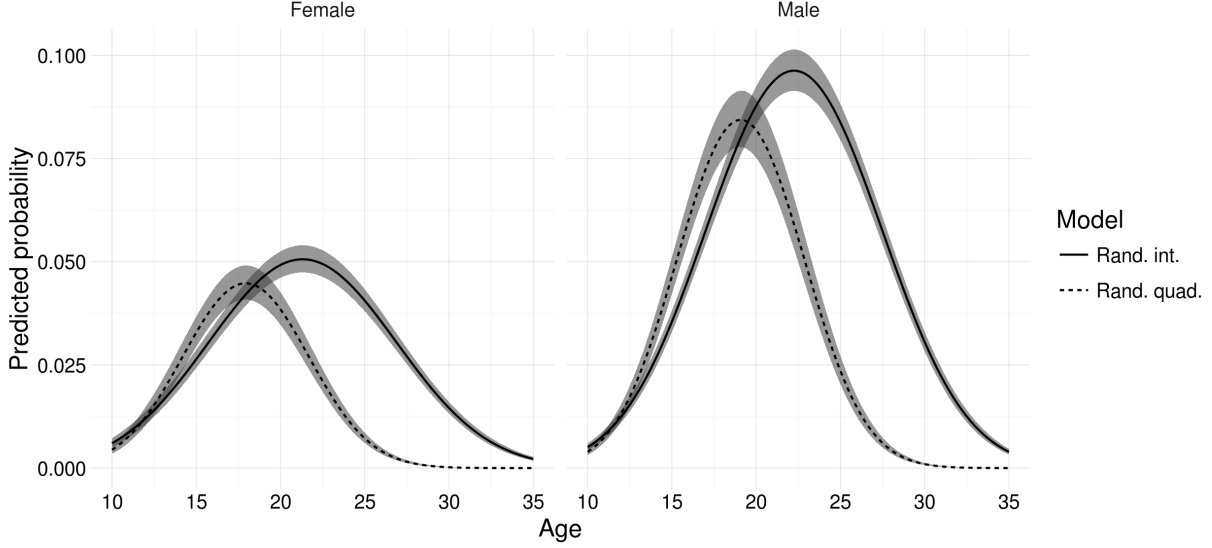


Figure 1: One-step correction point estimates and pointwise 95% confidence intervals of probability of having used marijuana in the past month. Curves on the left are for the average female, right are average male. Both the logistic regression with random intercepts and random quadratics are shown.

conditional probability  $p_{ij} = P(Y_{ij} = 1 | Z_i, SEX_i, AGE_{ij})$  we now have

$$\text{logit}(p_{ij}) = \begin{cases} (Z_{i0} + \beta_0) + (Z_{i1} + \beta_1)(AGE_{ij}/35) + (Z_{i2} + \beta_2)(AGE_{ij}/35)^2, & SEX_i = 0 \\ (Z_{i0} + \beta_3) + (Z_{i1} + \beta_4)(AGE_{ij}/35) + (Z_{i2} + \beta_5)(AGE_{ij}/35)^2, & SEX_i = 1. \end{cases}$$

Thus  $\beta_0, \beta_1$ , and  $\beta_2$  are the coefficients of the quadratic curve for the average female, and analogously for males. We model the random effects  $Z_n$  as IID mean zero multivariate Gaussian with covariance matrix  $\Sigma$ . Once again we use MLE, a Gaussian variational approximation, and a one-step correction to the Gaussian variational approximation to estimate the average random effects and covariance matrix.

We conduct a simulation study with the same structure as the study in the last section to compare the three methods. Box plots of the three estimators of the mean random effects are shown in Figure 2. The pattern is very different from the random intercept model. The `lme4` estimates are slightly inconsistent (for random effects with dimension larger than one the `lme4` package uses a Laplace approximation to the likelihood). The GVA estimates are even more biased than the `lme4` estimates. Despite this, it appears that our proposed

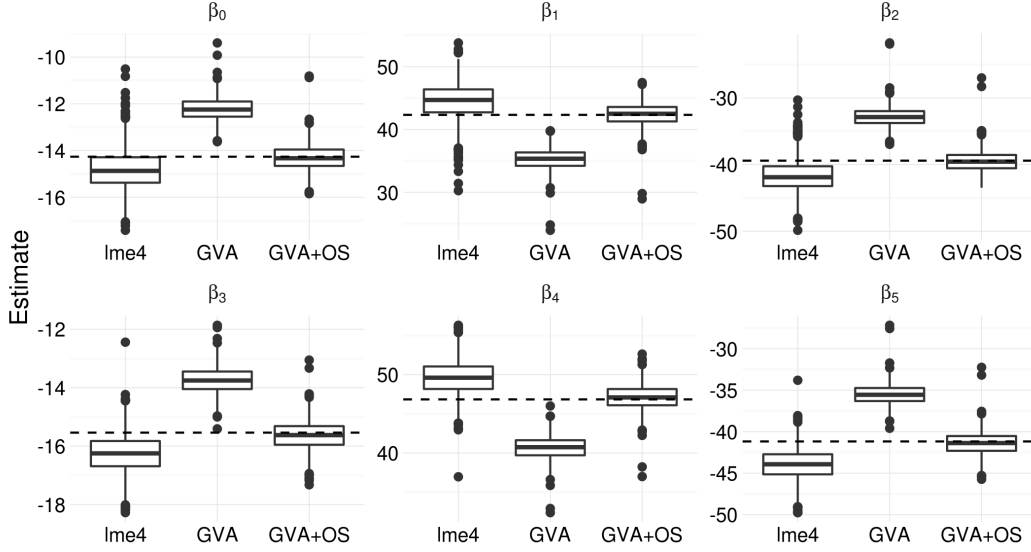


Figure 2: Estimates of mean random effects from the logistic regression with random quadratics simulation study. The dotted line indicates the true parameter value. “lme4” corresponds to estimate from the `lme4` package, which uses a Laplace approximation. “GVA” stands for Gaussian variational approximation, and “GVA+OS” refers to the one-step correction.

one-step corrected fixed effects are oughly centered around the truth. This is surprising since our theory does not guarantee that the one-step correction will be consistent when the variational estimate is not. All three estimators performed quite poorly in terms of estimating the covariance matrix of the random effects.

Table 4 shows the estimated coverage of 95% CIs for the mean random effects for the three estimators. The variational sandwich coverage was close to 0 in every case due to the bias in the parameter estimate seen in Figure 2. The `lme4` CIs also do not perform well, with substantially lower than desired coverage. The one-step correction coverage is closest to the desired 95%. These intervals are conservative, containing the true value more than 95% of the time.

The multivariate Hotelling test of consistency soundly rejected for every simulation, correctly indicating that the variational parameter estimator is consistent. In practice, therefore, while we would not be able to perform the same empirical evaluation as we have here (since we do not know the true model parameters to compute accuracy and coverage),

Table 4: Coverage of 95% confidence in the logistic regression with random quadratics simulation.

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$	$\beta_5$
Laplace approximation via <b>lme4</b>	0.72	0.68	0.60	0.67	0.61	0.52
GVA + sandwich	0.01	0.01	0.00	0.013	0.01	0.00
One-step correction to GVA	0.98	0.98	0.98	0.98	0.98	0.98

we would have information that calls into question the viability of the GVA procedure for this model. The marginal  $t$ -tests of consistency for  $\beta$  did not reject in the majority of simulations. Hence while the marginal  $t$ -tests were an accurate diagnostic tool in the random intercept setting, they were not in the random quadratic setting.

## 5 Discussion

We have presented a general framework for understanding the properties of variational estimation for parametric mixture models. The key insight of our work comes from representing the profiled variational objective function as an M-estimator. Once we make this connection, we can leverage a rich toolkit of asymptotic and methodological results available for this context.

The theory does not guarantee that variational estimators are consistent, and it is often difficult to derive the profile objective function necessary to assess consistency. We proposed an empirical test of consistency based on estimating the gradient of the profile objective at a single parameter value. This proposed method worked well in practice, correctly indicating whether variational estimator is inconsistent in two generalized linear mixed models.

We also used the asymptotic theory to propose a sandwich covariance estimator to provide calibrated confidence regions of variational estimators and a one-step correction to the variational estimator. Both of these methods work well when the variational estimator is consistent, and in fact the one-step correction exceeded our expectations by correcting some of the bias in fixed-effect variational parameter estimators in a logistic regression

model with random quadratics.

Our theory is limited to models which are IID at some level. While this includes many hierarchical and longitudinal models, it does exclude models for fully dependent time series, spatial data, and dyadic observations. Extending the theory to cover those cases could be a fruitful next step. Additionally we made the simplifying assumption that the variational class is of fixed and finite dimension, but our theory could be extended to more complicated variational classes.

## References

- Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. (2008), ‘Mixed membership stochastic blockmodels’, *Journal of Machine Learning Research* **9**, 1981–2014.
- Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015), ‘Fitting linear mixed-effects models using lme4’, *Journal of Statistical Software* **67**(1), 1–48.
- Bickel, P., Choi, D., Chang, X. & Zhang, H. (2013), ‘Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels’, *The Annals of Statistics* **41**(4), 1922–1943.
- Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), ‘Latent dirichlet allocation’, *Journal of Machine Learning Research* **3**, 993–1022.
- Bureau of Labor Statistics, U.S. Department of Labor (2013), ‘National longitudinal survey of youth 1997 cohort, 1997-2011 (rounds 1-15)’, Produced by the National Opinion Research Center, the University of Chicago and distributed by the Center for Human Resource Research, The Ohio State University, Columbus, OH.
- Davidian, M. & Giltinan, D. M. (1995), *Nonlinear Models for Repeated Measurement Data*, Vol. 62, CRC press.
- Erosheva, E., Fienberg, S. & Lafferty, J. (2004), ‘Mixed-membership models of scientific publications’, *Proceedings of the National Academy of Sciences* **101**, 5220–5227.

- Fabio, A., Tu, L.-C., Loeber, R. & Cohen, J. (2011), ‘Neighborhood socioeconomic disadvantage and the shape of the age–crime curve’, *American Journal of Public Health* **101**, S325–S332.
- Goldstein, H. (2011), *Multilevel Statistical Models*, Vol. 922, John Wiley & Sons.
- Hall, P., Humphreys, K. & Titterton, D. (2002), ‘On the adequacy of variational lower bound functions for likelihood-based inference in Markovian models with missing values’, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**(3), 549–564.
- Hall, P., Ormerod, J. T. & Wand, M. P. (2011), ‘Theory of Gaussian variational approximation for a Poisson mixed model’, *Statistica Sinica* **21**(1), 369–389.
- Hall, P., Pham, T., Wand, M. P. & Wang, S. S. (2011), ‘Asymptotic normality and valid inference for gaussian variational approximation’, *The Annals of Statistics* **39**(5), 2502–2532.
- Lancaster, T. (2000), ‘The incidental parameter problem since 1948’, *Journal of Econometrics* **95**(2), 391 – 413.
- Lee, C. Y. Y. & Wand, M. P. (2016), ‘Variational methods for fitting complex Bayesian mixed effects models to health data’, *Statistics in Medicine* **35**, 165–188.
- McCulloch, C. E. & Neuhaus, J. M. (2001), *Generalized Linear Mixed Models*, Wiley Online Library.
- O’Connor, B., Eisenstein, J., Xing, E. P. & Smith, N. A. (2010), ‘Discovering demographic language variation’, *Technical report*.
- Ormerod, J. T. & Wand, M. P. (2010), ‘Explaining variational approximations’, *The American Statistician* **64**(2), 140–153.
- Pritchard, J. K., Stephens, M. & Donnelly, P. (2000), ‘Inference of population structure using multilocus genotype data’, *Genetics* **155**(2), 945–959.

- Raj, A., Stephens, M. & Pritchard, J. K. (2014), ‘fastSTRUCTURE: Variational inference of population structure in large snp data sets’, *Genetics* **197**(2), 573–589.
- van der Vaart, A. W. (2000), *Asymptotic Statistics*, Vol. 3, Cambridge University Press.
- Wainwright, M. J. & Jordan, M. I. (2008), ‘Graphical models, exponential families, and variational inference’, *Foundations and Trends in Machine Learning* **1**(1-2), 1–305.
- Wang, B. & Titterton, D. M. (2004), Convergence and asymptotic normality of variational Bayesian approximations for exponential family models with missing values, *in* ‘Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence’, UAI ’04, AUAI Press, Arlington, Virginia, United States, pp. 577–584.
- Wang, B. & Titterton, D. M. (2006), ‘Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model’, *Bayesian Analysis* **1**(3), 625–650.

# Appendices

## A Example of underestimated uncertainty for variational approximations

As discussed in the main text, even when the variational Bayes posterior is consistent, it frequently underestimates the true posterior variance. This phenomenon can be explained intuitively using the KL divergence between multivariate normal distributions. Under regularity conditions, the posterior distribution of model parameters  $\Pi_n(\theta|X_{1:n})$  looks approximately  $N_D(\theta_0, \Sigma)$  as  $n$  grows (where  $\Sigma$  implicitly depends on  $n$  because  $\theta$  has not been appropriately rescaled). Often the variational distribution is asymptotically normal as well. However, if the variational class of distributions over model parameters only includes factored distributions, then the variational distribution can only be approaching independent normal distributions. The KL divergence between a normal distribution with mean  $\mu$  and diagonal covariance matrix with  $k$ th diagonal entry  $\sigma_k^2$  and a general multivariate normal is minimized when  $\mu = \theta_0$  and  $\sigma_k^2 = 1/(\Sigma^{-1})_{kk}$ . However, using Schur complements we can see that

$$\sigma_k^2 = \frac{1}{(\Sigma^{-1})_{kk}} = \Sigma_{kk} - \Sigma_{k\cdot}(\Sigma_{-kk})^{-1}\Sigma_{k\cdot}^T,$$

where  $\Sigma_{k\cdot}$  is the  $k$ th row of  $\Sigma$  omitting  $\Sigma_{kk}$  and  $\Sigma_{-kk}$  is the minor of  $\Sigma$  removing the  $k$ th row and  $k$ th column. Assuming  $\Sigma$  is positive definite,  $\Sigma_{-kk}^{-1}$  is positive definite as well and hence  $\Sigma_{k\cdot}\Sigma_{-kk}^{-1}\Sigma_{k\cdot}^T \geq 0$  with equality if and only if  $\Sigma_{k\cdot} = \mathbf{0}$ . Hence  $\sigma_k^2$ , the marginal variational posterior variance of  $\theta_k$ , is  $\leq \Sigma_{kk}$ , the true marginal posterior variance, with equality if and only if  $\theta_k$  is not correlated in the posterior with any of the other model parameters. Thus, we should expect the variational Bayes posterior to underestimate the marginal uncertainty of any model parameter or latent variable that is asymptotically correlated with other model parameters or latent variable.



## B Proof of theorems

Recall that  $P_0$  is the true distribution,  $\mathcal{Q}$  is the variational family of distributions over the latent variable  $Z$ , which is parametrized by  $\psi \in \Psi$ , and

$$v(\theta, \psi; x) = E_\psi \left[ \log \frac{p_\theta(x, Z)}{q(Z; \psi)} \right]$$

is one term in the variational criterion function.

We start with a list of assumptions we will need for consistency:

**(A1)** The map  $\theta \mapsto \sup_{\psi \in \Psi} v(\theta, \psi; x)$  is upper-semicontinuous a.s.- $P_0$ .

**(A2)** There exists a  $d > 0$  such that for all  $\delta < d$  and  $\eta \in \Theta$  the map

$$x \mapsto \sup_{\substack{\theta \in B_\delta(\eta) \\ \psi \in \Psi}} v(\theta, \psi; x)$$

is measurable and

$$E_{P_0} \sup_{\substack{\theta \in B_\delta(\eta) \\ \psi \in \Psi}} v(\theta, \psi; X) < \infty.$$

**(A3)** There exists a compact set  $K \subset \Theta$  such that  $P_0(\hat{\theta}_n \in K) \rightarrow 1$ .

We can now demonstrate Theorem 1.

*Proof of Theorem 1.* Defining  $m(\theta; x) = \sup_{\psi \in \Psi} v(\theta, \psi; x)$ , the requirements of Theorem 5.14 of van der Vaart (2000) are satisfied. Hence for all  $\epsilon > 0$ ,  $P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) \rightarrow 0$ . Since  $P_0(\hat{\theta}_n \in K) \rightarrow 1$  by assumption,

$$\begin{aligned} P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon) &\leq P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) + P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K^c) \\ &\leq P_0(\|\hat{\theta}_n - \bar{\theta}\| \geq \epsilon \cap \hat{\theta}_n \in K) + P_0(\hat{\theta}_n \in K^c) \rightarrow 0. \end{aligned}$$

□

We next lay out sufficient conditions for asymptotic normality. Throughout we will assume that for all  $\theta$ ,  $(\psi, x) \mapsto v(\theta, \psi; x)$  is a measurable function on the product measure space  $\Psi \times \mathcal{X}$ , where  $\Psi$  is equipped with Borel measure.

- (B1) For all  $\theta$  and  $P_0$ -a.e.  $x$ ,  $v(\theta, \psi; x)$  is uniquely maximized at  $\psi^*(\theta; x)$  which is an element of  $\Psi$ , an open subset of  $\mathbb{R}^d$ .
- (B2)  $\psi^*$  is a measurable function of  $x$  for all  $\theta$  and twice continuously differentiable in a neighborhood of  $\bar{\theta}$  for  $P_0$ -a.e.  $x$ .
- (B3)  $v$  is twice continuously differentiable in a neighborhood of  $\bar{\theta}$  and  $\psi^*(\bar{\theta}; x)$  for  $P_0$ -a.e.  $x$ .
- (B4) There exist  $r > 0$ ,  $s(x) > 0$ ,  $b_1(x)$  and  $b_2(x)$  such that
- (a) For all  $x \in \mathcal{X}$  and  $\theta \in \mathcal{B}_r(\bar{\theta})$ ,  $\psi^*(\theta; x) \in \mathcal{B}_{s(x)}(\psi^*(\bar{\theta}; x))$
  - (b) For all  $x \in \mathcal{X}$ ,  $\theta_1, \theta_2 \in \mathcal{B}_r(\bar{\theta})$  and  $\psi_1, \psi_2 \in \mathcal{B}_{s(x)}(\psi^*(\bar{\theta}; x))$ ,
- $$|v(\theta_1, \psi_1; x) - v(\theta_2, \psi_2; x)| \leq b_1(x)(\|\theta_1 - \theta_2\| + \|\psi_1 - \psi_2\|).$$
- (c) For all  $\theta_1, \theta_2 \in \mathcal{B}_r(\bar{\theta})$ ,  $\|\psi^*(\theta_1; x) - \psi^*(\theta_2; x)\| \leq b_2(x)\|\theta_1 - \theta_2\|$ .
  - (d)  $b_1$  and  $b_1 b_2 \in L_2(P_0)$ .
- (B5)  $|D_{\theta}^2 v(\theta, \psi^*(\theta; x); x)| \leq \kappa(x)$  for all  $\theta$  in a neighborhood of  $\bar{\theta}$  and  $P_0$ -a.e.  $x$  for an integrable function  $\kappa$ .

With these conditions we prove Theorem 2.

*Proof of Theorem 2.* We will use van der Vaart (2000) Theorem 5.23. We need to validate the following conditions to apply the result: 1.  $m(\theta; x)$  is measurable as a function of  $x$  for all  $\theta \in \Theta$ ; 2.  $m(\theta; x)$  is differentiable at  $\bar{\theta}$  for  $P_0$ -a.e.  $x$ ; 3. there exists a measurable function  $b \in L_2(P_0)$  and an  $r > 0$  such that for all  $\theta_1, \theta_2 \in \mathcal{B}_r(\bar{\theta})$ ,  $|m(\theta_1; x) - m(\theta_2; x)| \leq b(x)\|\theta_1 - \theta_2\|$ ; 4. The function  $m(\theta) = E_{P_0}[m(\theta; X)]$  is maximized at  $\theta = \bar{\theta}$  and admits a second-order Taylor expansion at  $\bar{\theta}$ ; 5.  $\frac{1}{n} \sum_i m(\hat{\theta}_n; X_i) \geq \sup_{\theta \in \Theta} \frac{1}{n} \sum_i m(\hat{\theta}; X_i) - o_P(1)$ . We will demonstrate that these conditions follow from assumptions (B1)-(B5).

For the first condition, the measurability of  $x \mapsto m(\theta; x)$  is guaranteed by the measurability of  $\psi^*$  and  $v$  plus the fact that compositions of measurable functions are measurable.

Differentiability of  $m$  at  $\bar{\theta}$  is implied by conditions (B2) and (B3) together with the multivariate chain rule. We have  $(D_{\theta} m)(\bar{\theta}; x) = (D_{\theta} v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) + (D_{\theta} \psi^*)(\bar{\theta}; x)^T (D_{\psi} v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x)$ .

Since  $\psi \mapsto v(\bar{\theta}, \psi; x)$  is maximized at  $\psi^*(\bar{\theta}; x)$ , which is in the interior of  $\Psi$ , and  $v$  is differentiable in  $\psi$  at  $\bar{\theta}, \psi^*(\bar{\theta}; x)$  for  $P_0$ -a.e.  $x$ ,  $(D_\psi v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) = 0$  a.s.  $P_0$ .

For the third condition we apply (B3). Let  $\theta_1, \theta_2 \in \mathcal{B}_r(\bar{\theta})$ . Then for each  $x$   $\psi^*(\theta_1; x), \psi^*(\theta_2; x) \in \mathcal{B}_{s(x)}(\psi^*(\bar{\theta}; x))$ . Hence

$$\begin{aligned} |m(\theta_1; x) - m(\theta_2; x)| &= |v(\gamma^*(\theta_1; x); x) - v(\gamma^*(\theta_2; x); x)| \leq b_1(x) (\|\theta_1 - \theta_2\| + \|\psi^*(\theta_1; x) - \psi^*(\theta_2; x)\|) \\ &\leq b_1(x) (\|\theta_1 - \theta_2\| + b_2(x)\|\theta_1 - \theta_2\|) = b_1(x)(1 + b_2(x))\|\theta_1 - \theta_2\|. \end{aligned}$$

Since by assumption  $b_1, b_1 b_2 \in L_2(P_0)$ , the third condition is satisfied with  $b = b_1(1 + b_2)$ .

Assumptions (B2), (B4) and (B5) imply that the map  $\theta \mapsto E_{P_0}[v(\theta, \psi^*(\theta; x); x)] = E_{P_0}[m(\theta; x)]$  is twice continuously differentiable in a neighborhood of  $\bar{\theta}$  and hence possesses a second-order Taylor expansion, thus satisfying the fourth condition above. Furthermore, these assumptions justify differentiation under the integral. We can thus derive the second derivative matrix of  $m(\theta; x)$  at  $\bar{\theta}$  as follows:

$$D_\theta^2 m(\theta; x) = D_\theta(D_\theta m(\theta; x)) = D_\theta(D_\theta v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) \quad (7)$$

$$= (D_\theta^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) + (D_{\theta\psi}^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x)(D_\theta \psi^*)(\bar{\theta}; x) \quad (8)$$

By definition  $\psi^*(\bar{\theta}; x)$  solves  $(D_\psi v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) = 0$ . Differentiating with respect to  $\theta$  gives

$$0 = (D_{\theta\psi}^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) + (D_\theta \psi^*)(\bar{\theta}; x)(D_{\psi\psi}^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x) \quad (9)$$

$$(D_\theta \psi^*)(\bar{\theta}; x) = -(D_{\theta\psi}^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x)(D_{\psi\psi}^2 v)(\bar{\theta}, \psi^*(\bar{\theta}; x); x)^{-1}. \quad (10)$$

Solving for  $(D_\theta \psi^*)(\bar{\theta}; x)$  and substituting this back in to (8) gives

$$D_{\theta\theta}^2 m(\theta; x) = (D_{\theta\theta}^2 v - (D_{\theta\psi}^2 v)(D_{\psi\psi}^2 v)^{-1}(D_{\theta\psi}^2 v)^T)(\bar{\theta}, \psi^*(\bar{\theta}; x); x). \quad (11)$$

All the derivatives are now in terms of  $v$ , a known function, and evaluated at  $\bar{\theta}, \psi^*(\bar{\theta}; x)$ .

Finally, condition five is satisfied since  $\hat{\theta}_n$  maximizes  $\frac{1}{n} \sum_i m(\hat{\theta}; X_i)$  by definition. This establishes the asymptotic normality.  $\square$

## C Analysis of exponential mixture model

Recall the exponential mixture model presented in the main text: we observe pairs  $(X_{i,1}, X_{i,2})$  for  $i = 1, \dots, n$  generated independently and identically with  $X_{i,1} \sim \text{Exp}(\lambda_1)$ , and  $X_{i,2}|Z_i \sim \text{Exp}(\lambda_1 Z_i)$  where  $Z_i$  is a latent variable with  $Z_i \sim \text{Exp}(\lambda_2)$ .

Suppose we use variational inference where the variational class is taken to be all log-normal distributions parametrized by  $\psi = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}^+ = \Psi$ . Here we evaluate the properties of the resulting variational estimators.

The variational criterion function is

$$\begin{aligned} v(\lambda, \psi; x) &= 2 \log \lambda_1 - \lambda_1 x_1 + \log \lambda_2 + E_\psi[\log Z] - (\lambda_1 x_2 + \lambda_1) E_\psi[Z] - E_\psi[\log q(Z; \psi)] \\ &= 2 \log \lambda_1 - \lambda_1 x_1 + \log \lambda_2 + 2\mu - (\lambda_1 x_2 + \lambda_1) e^{\mu + \sigma^2/2} + \log \sigma + 1/2. \end{aligned}$$

Differentiating with respect to  $\psi$  and setting to zero we get the closed-form solutions  $\hat{\sigma} = 1/\sqrt{2}$  and  $\hat{\mu} = -\frac{1}{4} + \log\left(\frac{2}{\lambda_1 x_2 + \lambda_1}\right)$ . Plugging this back in to the criterion gives

$$v(\lambda, \hat{\psi}(\lambda; x); x) = 2 \log \lambda_1 - \lambda_2 x_1 + \log \lambda_2 - 2 \log(\lambda_1 x_2 + \lambda_1) + c.$$

This is the marginal log-likelihood of a single observation plus a constant. Hence the variational estimates of  $\lambda$  using this variational class would be exactly equal both to the mean-field and maximum likelihood estimates.

## D Additional simulation results

In the main text we tabulated the variances of the three estimators for the random intercepts simulation but did not show the raw estimates. Figure 3 contains box plots of the estimators of each of the seven parameters (the average random effects for females and males  $\beta_0$  and  $\beta_3$ , the fixed linear and quadratic effects of age for females,  $\beta_1$  and  $\beta_2$ , and for males,  $\beta_4$  and  $\beta_5$ , and the log variance of the random effects  $\log(\sigma^2)$ ).

As discussed in the main text, all three methods appear to be consistent for all elements of  $\beta$ , but only **lme4** is consistent for  $\log(\sigma^2)$ . The raw variational estimates of  $\beta$  are slightly less efficient than those of **lme4** or the one-step correction.

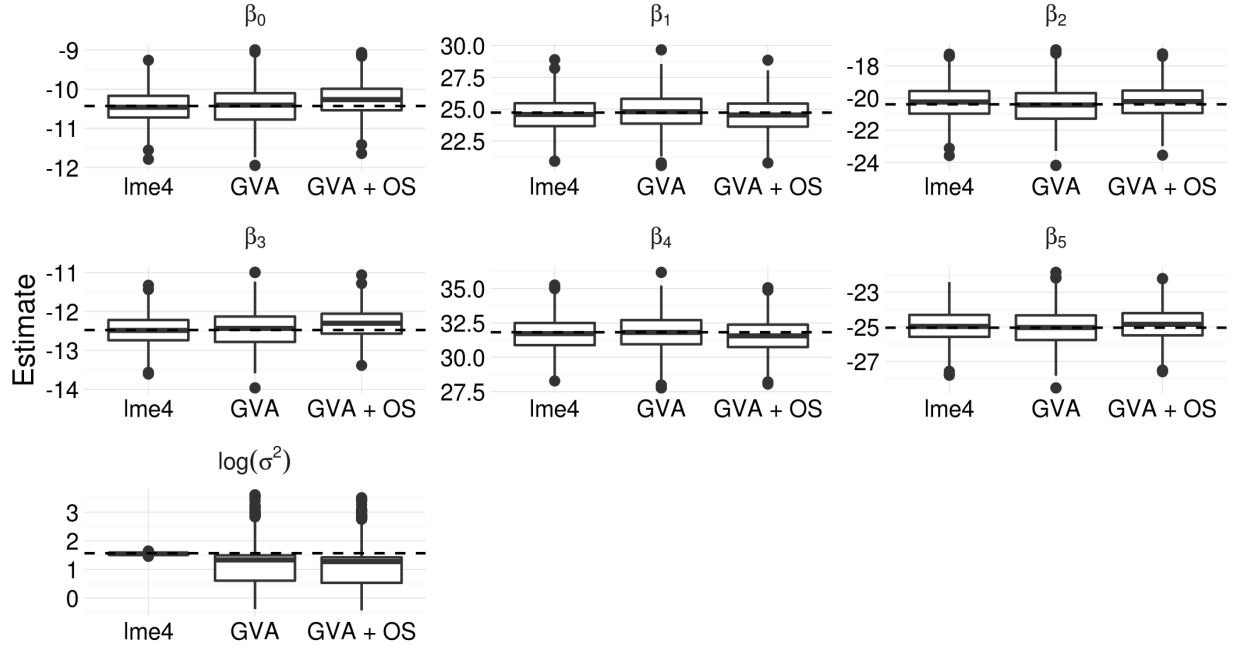


Figure 3: Boxplots of parameter estimates from the logistic regression with random intercepts simulation study. “lme4” corresponds to estimate from the `lme4` package, “GVA” stands for Gaussian variational approximation, and “GVA+OS” refers to the one-step correction.