The Annals of Applied Statistics
2018, Vol. 12, No. 2, 846–876
https://doi.org/10.1214/18-AOAS1169

ON THE USE OF BOOTSTRAP WITH VARIATIONAL INFERENCE: THEORY, INTERPRETATION, AND A TWO-SAMPLE TEST EXAMPLE

BY YEN-CHI CHEN, Y. SAMUEL WANG AND ELENA A. EROSHEVA

University of Washington

Variational inference is a general approach for approximating complex density functions, such as those arising in latent variable models, popular in machine learning. It has been applied to approximate the maximum likelihood estimator and to carry out Bayesian inference, however, quantification of uncertainty with variational inference remains challenging from both theoretical and practical perspectives. This paper is concerned with developing uncertainty measures for variational inference by using bootstrap procedures. We first develop two general bootstrap approaches for assessing the uncertainty of a variational estimate and the study the underlying bootstrap theory in both fixed- and increasing-dimension settings. We then use the bootstrap approach and our theoretical results in the context of mixed membership modeling with multivariate binary data on functional disability from the National Long Term Care Survey. We carry out a two-sample approach to test for changes in the repeated measures of functional disability for the subset of individuals present in 1989 and 1994 waves.

1. Introduction. Variational inference [Jordan et al. (1999), Wainright and Jordan (2008)] is a method to approximate complex density functions [Blei, Kucukelbir and McAuliffe (2017)] which has been applied to various statistical models such as factor analysis [Ghahramani and Beal (2000), Khan et al. (2010), Klami et al. (2015)], stochastic block models [Celisse, Daudin and Pierre (2012), Latouche, Birmelé and Ambroise (2012), Bickel et al. (2013)], latent Dirichlet allocation [Blei and Jordan (2006), Blei, Ng and Jordan (2003)], and Gaussian processes [Damianou, Titsias and Lawrence (2011, 2016)].

Variational inference can be used to approximate a posterior distribution as an alternative to Markov Chain Monte Carlo (MCMC), when a sampling procedure would be prohibitively slow or require immense human effort to tune, or to approximate a maximum likelihood estimator (MLE), when computation with the specified likelihood is intractable. In particular, when the model involves a latent structure such as a mixed membership model [Airoldi et al. (2005, 2008), Wang, Matsueda and Erosheva (2017)] or a mixed effect model [Hall, Ormerod and Wand (2011a), Westling and McCormick (2015)], finding the MLE may be very challenging and variational inference provides a fast way to obtain an estimate of the

Received November 2017; revised April 2018.

Key words and phrases. Variational inference, bootstrap, mixed membership model, increasing dimension, two-sample test.

parameter. The estimator from variational inference is called the variational estimator.

Recently, the asymptotic distribution of point estimates resulting from variational inference was investigated in Bickel et al. (2013), Hall et al. (2011b), Westling and McCormick (2015) and Wang and Blei (2017) analyze variational inference under a Bayesian framework. When a consistent estimator of the asymptotic variance is available, practitioners can analyze the uncertainty of the variational estimate and draw scientific conclusions by constructing confidence intervals (CI) for the parameters of interest [Hall et al. (2011b), Westling and McCormick (2015)].

However, constructing a CI using the asymptotic distribution fails if we do not have a consistent estimator of the variance of the variational estimator. To overcome this problem, we consider using bootstrap methods implemented in Bickel et al. (2013) and Wang, Matsueda and Erosheva (2017). The bootstrap approach does not require a consistent variance estimator to be available, and, in some cases, leads to a CI with a higher-order coverage [Hall (1992)]. Despite the fact that the bootstrap method has already been used with variational estimation [Wang, Matsueda and Erosheva (2017)], the underlying bootstrap theory for variational inference does not exist.

In this paper, we investigate the validity of using a bootstrap approach where variational inference is used to approximate an MLE. We construct a confidence interval (CI) in the usual fixed dimensional case, where both the dimensionality of the parameter and the number of latent variables are fixed, as well as in the increasing dimensional case. An example of the latter situation may come from an item response theory model where the latent dimensionality may increase when the number of questions per individual is increasing. Haberman (1977) and Douglas (1997) have analyzed a situation where the number of questions (dimension) and the number of participants (sample size) increase jointly.

This paper has been motivated by the general need—as opposed to one specific substantive problem or a specific application area—to provide statisticians, computer scientists, and data scientists with the theory and tools for using the bootstrap for variational inference. We use two sets of functional disability measures obtained five years apart from the National Long Term Care Survey (NLTCS) to illustrate the bootstrap approach on a two-sample test, a setting where we find the variational inference to be particularly appropriate. However, a complete development of a substantive application is beyond the scope of this paper.

Outline. We briefly review variational inference in Section 2. In Section 3, we discuss how to apply the bootstrap to variational inference. We then develop asymptotic normality and bootstrap theory in Section 4. In Section 5, we illustrate the bootstrap approach with a two-sample test using functional disability data from the NLTCS. Finally, we discuss related topics and the link to Stephen E. Fienberg in Section 6.

2. Variational inference. We consider the variational inference in the context of a latent variable model. Assume our data consists of n individuals and J variables (e.g., survey questions or test problems) and forms a random sample of $X_1, \ldots, X_n \in \mathbb{R}^J$ that are IID from a distribution function P_0. We assume that there exists K latent features for each individual that are denoted as $Z_1, \ldots, Z_n \in \mathbb{R}^K$. This setup is quite general—in a mixed membership model, Z_i is the vector of mixed membership indicator; in a random effect model, Z_i is the random effect; in a stochastic block model, Z_i is the community indicator (and $X_i = \{0, 1\}^n$ denotes the edge connected to the ith observation).

We assume a parametric model on the distribution function P_0 such that the joint distribution of (X_i, Z_i) has a parametric form $P(x, z; \theta)$, where $\theta \in \Theta \subset \mathbb{R}^d$ is the parameter of interest. When the latent feature vector Z_i is known, the likelihood of ith observation is

$$L(\theta|X_i, Z_i) = P(X_i, Z_i; \theta).$$

Analogously to Neyman and Scott (1948), we regard the latent feature vectors $Z_1, \ldots, Z_n$ as incidental parameters and the population parameter θ as structural parameters.

In reality, we do not know the latent vectors so the observed log-likelihood function is

$$\ell(\theta|\mathcal{X}) = \log L(\theta|X_i) = \log \int P(X_i, Z_i; \theta)\, dZ_i. \tag{1}$$

Often, we are interested in using the maximum likelihood estimator

$$\hat{\theta}_{\mathrm{MLE}} = \arg\max_{\theta} \sum_{i=1}^{n} \ell(\theta|X_i).$$

However, maximizing or even calculating the marginal likelihood can often be computationally intractable. Thus, variational estimators provide an alternative, computationally feasible estimator to the MLE. The variational estimator is constructed as follows. We first pick a family of distributions —the variational distribution family—for the latent variable Z_i. Let $Q(z; \omega)$ be the variational distribution family indexed by the variational parameter $\omega \in \Omega \subset \mathbb{R}^s$, which is a nuisance parameter in our model. Note that we allow each Z_i has its own variational parameter; namely, $\omega = \omega_i$. Using Jensen's inequality, the log-likelihood function satisfies

$$\begin{aligned}
\ell(\theta|X_i) &= \log \int P(X_i, Z_i; \theta)\, dZ_i \\
&= \log \int \frac{P(X_i, Z_i; \theta)}{Q(Z_i; \omega_i)} Q(Z_i; \omega_i)\, dZ_i \\
&= \log \mathbb{E}_{Z_i \sim Q}\left(\frac{P(X_i, Z_i; \theta)}{Q(Z_i; \omega_i)} \middle| X_i \right) \qquad (2)
\end{aligned}$$

$$\geq \mathbb{E}_{Z_i \sim Q}\big(\log P(X_i, Z_i; \theta)|X_i\big) - \mathbb{E}_{Z_i \sim Q}\big(\log Q(Z_i; \omega_i)\big)$$
$$= \mathsf{ELBO}(\theta, \omega_i|X_i),$$

where $\mathbb{E}_{Z_i \sim Q_i}$ means that the expectation is taken over variable Z_i and the underlying distribution is $Q(\cdot; \omega_i)$. We call the expression on the right hand side of the inequality the *evidence lower bound* (ELBO).

Instead of maximizing the log-likelihood function, the variational framework maximizes the ELBO, leading to

$$\hat{\theta}_{\mathrm{ELBO}}, \hat{\omega}_{\mathrm{ELBO},1}, \ldots, \hat{\omega}_{\mathrm{ELBO},n} = \operatorname*{arg\,max}_{\theta, \omega_1, \ldots, \omega_n} \sum_{i=1}^{n} \mathsf{ELBO}(\theta, \omega_i|X_i). \tag{3}$$

Because ω_i in the above maximizing criterion is only involved in $\mathsf{ELBO}(\theta, \omega_i|X_i)$ when θ is fixed, the first element $\hat{\theta}_{\mathrm{ELBO}}$ is equivalent to the maximizer of the following criterion:

$$\begin{aligned}\hat{\theta}_{\mathrm{ELBO}} &= \operatorname*{arg\,max}_{\theta} \sum_{i=1}^{n} \mathsf{ELBO}\big(\theta, \omega_{\max}(\theta|X_i)|X_i\big) \\ &= \operatorname*{arg\,max}_{\theta} \sum_{i=1}^{n} \mathcal{E}(\theta|X_i),\end{aligned} \tag{4}$$

where $\omega_{\max}(\theta|X_i) = \operatorname{arg\,max}_{\omega_i} \mathsf{ELBO}(\theta, \omega_i|X_i)$. The quantity $\hat{\theta}_{\mathrm{ELBO}}$, is called the ELBO estimator or the variational estimator.

Because the ELBO estimator comes from optimizing $\sum_{i=1}^{n} \mathcal{E}(\theta|X_i)$, it is an estimator of

$$\theta_{\mathrm{ELBO}} = \operatorname*{arg\,max}_{\theta} \mathbb{E}\big(\mathcal{E}(\theta|X_1)\big). \tag{5}$$

Note that the expectation in the above expression is for the random variable X_1 and is taken with respect to the data-generating distribution P_0. The quantity θ_{ELBO} defines the population quantity that the variational inference (ELBO estimator) is estimating. Note that θ_{ELBO} depends on the variational distribution Q and is often different from the population version of $\theta_{\mathrm{MLE}} = \operatorname{arg\,max}_{\theta} \mathbb{E}(\ell(\theta|X_1))$. Thus, variational inference can be thought of as an intentional model misspecification even if the original parametric model is correctly specified. We will argue in the next section that despite the misspecification, variational inference is still a useful procedure for making statistical inference.

REMARK 1. When the parametric model is correctly specified (i.e., there exists $\theta_0 \in \Theta$ such that $P_0 = P_{\theta_0}$), the variational estimator may recover the correct model with $\theta_{\mathrm{ELBO}} = \theta_0$ in some special cases. For concrete examples, we refer the readers to Hall et al. (2011b) and Bickel et al. (2013) where they illustrated this possibility in a single predictor Poisson mixed model and a stochastic block model.

2.1. *Further considerations for using variational inference in practice.* As described in the previous section, the distribution based on the variational estimator $P_{\hat{\theta}_{\text{ELBO}}}$ may not converge to the true data-generating distribution function even when the model is correctly specified. Despite this drawback, variational inference can be a useful procedure for inference for the following reasons.

- *Likelihood formulation as a working model.* As George Box has said *"Essentially, all models are wrong, but some are useful"* [Box (1976)]. A proposed model is almost always misspecified. When using a parametric model to analyze the data, we do not claim that the parametric model describes the actual data-generating distribution. Instead, a working model and parameter estimates help us to learn various aspects about the data at hand. To carry this reasoning further, the ML procedure and the variational inference procedure are just different principles of fitting parameters to the data. When the model is misspecified, both the MLE and the variational estimator are best approximation estimators under different criteria of measuring the quality of approximation. Figure 1 provides a diagram illustrating the case where the likelihood model does not include the true data-generating distribution.
- *Two-sample test.* The variational inference procedure is a useful procedure for two-sample test. Given two sets of data $X_1, \ldots, X_n \sim P_X$ and $Y_1, \ldots, Y_m \sim P_Y$, the goal of a two-sample test is to test

$$H_0 : P_X = P_Y.$$

Using the equation (5), H_0 implies that

$$\theta_{\text{ELBO},X} = \theta_{\text{ELBO},Y},$$

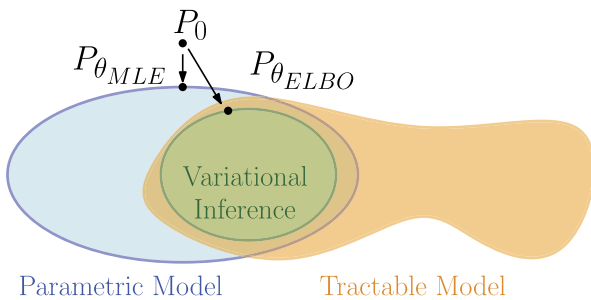

FIG. 1. *An illustration for the relations of* P_0, $P_{\theta_{\text{MLE}}}$, *and* $P_{\theta_{\text{ELBO}}}$ *when a latent variable model is used and the model is not correctly specified. In this case, the distribution corresponding to the population MLE is just the distribution in the parametric family that minimizes the KL-divergence to the true distribution function. So* $P_{\theta_{\text{MLE}}}$ *can be viewed as a projection from* P_0 *onto the parametric family. However, if the MLE is computationally intractable, we can still specify a tractable variational estimator and the corresponding variational distribution,* $P_{\theta_{\text{ELBO}}}$, *can be viewed as another projection from* P_0.

where $\theta_{\mathrm{ELBO},X}$ and $\theta_{\mathrm{ELBO},Y}$ are the maximizer of equation (5) assuming the expectation is taken over P_X and P_Y. Thus, when applied to a two-sample test, variational inference is as valid of an approach as ML inference. In a sense, one can interpret the tests following either approach, variational or ML, as inferences based on different projections of the distributions P_X, P_Y onto the same parameter space.

3. Bootstrapping the variational estimator. We use the bootstrap [Efron (1982a, 1982b)] to evaluate the uncertainty of the variational estimator and construct CIs. We focus on the empirical bootstrap—also known as classical, nonparametric, or Efron's bootstrap—where one samples with replacement from the original dataset, recomputes the ELBO estimator for each bootstrap sample, and uses the distribution of these bootstrapped ELBO estimators to derive uncertainty measures. We illustrate estimation of the error of $\hat{\theta}_{\mathrm{ELBO}}$ and construction of the CIs using the bootstrap. There are many bootstrap CIs [see, e.g., Hall (1992)]. Here, we will focus on two common approaches: the percentile method and the (studentized) pivotal method. Note that constructing a CI using the percentile approach has been implemented in Wang, Matsueda and Erosheva (2017).

The bootstrap approach to estimating uncertainty is very general. The bootstrap percentile approach can be used even when the asymptotic covariance matrix is not available (e.g., difficult to estimate). When the asymptotic covariance matrix of $\hat{\theta}_{\mathrm{ELBO}}$ is known and can be consistently estimated (say using a sandwich estimator), the bootstrap pivotal method may produce CIs with a higher order coverage than those based on the asymptotic normality [Babu and Singh (1983), Horowitz (1997), Singh (1981)].

More formally, let $X_1^*, \ldots, X_n^*$ be a bootstrap sample from the original sample $\mathcal{X} = \{X_1, \ldots, X_n\}$. Given the bootstrap sample, we compute the bootstrap ELBO estimator

$$\hat{\theta}^*_{\mathrm{ELBO}} = \arg\max_{\theta} \sum_{i=1}^{n} \mathcal{E}(\theta | X_i^*). \tag{6}$$

Repeating the bootstrap procedure B times, we obtain B bootstrap ELBO estimators:

$$\hat{\theta}^{*(1)}_{\mathrm{ELBO}}, \ldots, \hat{\theta}^{*(B)}_{\mathrm{ELBO}}.$$

We will use these bootstrap ELBO estimators to assess the uncertainty of the original ELBO estimator.

Note that one may also use the jackknife and weighted bootstrap [O'Hagan, Murphy and Gormley (2015)] to generate bootstrap sample. In particular, when analyzing a network data where each X_i corresponds to the edges of ith vertex, the empirical bootstrap cannot be applied but the weighted bootstrap (and the parametric bootstrap; see Remark 2) is still applicable.

3.1. *Estimating the variance*. The bootstrap approach can be applied to estimate the variance of the ELBO estimator. Assume that we focus on the ℓth parameter θ_ℓ. The variance of $\hat{\theta}_{\mathrm{ELBO},\ell}$ can be estimated using the sample variance of the bootstrapped variational estimators

$$\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{ELBO},\ell}) = \frac{1}{B}\sum_{j=1}^{B}\big(\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell} - \bar{\theta}^{*}_{\mathrm{ELBO},\ell}\big)^2, \qquad \bar{\theta}^{*}_{\mathrm{ELBO},\ell} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell}. \tag{7}$$

Figure 2 provides a diagram summarizing the procedure.

The intuition behind equation (7) is that the bootstrap distribution of the estimators $\hat{\theta}^{*(1)}_{\mathrm{ELBO}}, \ldots, \hat{\theta}^{*(B)}_{\mathrm{ELBO}}$ behaves as if new realizations of the original estimator $\hat{\theta}_{\mathrm{ELBO}}$ are drawn. Thus, the variance of the bootstrap estimators would be an approximation to the variance of $\hat{\theta}_{\mathrm{ELBO}}$.

3.2. *Confidence interval*: *Percentile approach*. The bootstrap approach enables us to construct CIs for the parameters of interest. We first introduce a simple approach called the percentile (quantile) approach, which is based on the percentile of the distribution of the bootstrap variational estimators. Assume again we focus on the ℓth parameter. Given a confidence level α, let $\hat{s}_{\ell,\alpha}$ denotes the α-quantile of the bootstrap ELBO estimators

$$\hat{s}_{\ell,\alpha} = \hat{G}_\ell^{-1}(1-\alpha), \qquad \hat{G}_\ell(t) = \frac{1}{B}\sum_{j=1}^{B} I\big(\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell} - \hat{\theta}_{\mathrm{ELBO},\ell} \le t\big).$$

Then a $(1-\alpha)$ CI of the ℓth parameter is

$$C_{n,\alpha,\ell} = [\hat{\theta}_{\mathrm{ELBO},\ell} + \hat{s}_{\ell,\alpha/2}, \hat{\theta}_{\mathrm{ELBO},\ell} + \hat{s}_{\ell,1-\alpha/2}]. \tag{8}$$

Figure 3 summarizes the steps in computing a bootstrap percentile CI.

BOOTSTRAP VARIANCE ESTIMATOR.

1. Find the variational estimator $\hat{\theta}_{\mathrm{ELBO}}$ using $X_1, \ldots, X_n$.
2. For $j = 1, \ldots, B$, do the following task:

(a) Sample with replacement from $X_1, \ldots, X_n$ to obtain $X_1^{*(j)}, \ldots, X_n^{*(j)}$.
(b) Compute the variational estimator $\hat{\theta}^{*(j)}_{\mathrm{ELBO}}$ using $X_1^{*(j)}, \ldots, X_n^{*(j)}$.

3. For the ℓth parameter, compute its variance estimator using

$$\widehat{\mathrm{Var}}(\hat{\theta}_{\mathrm{ELBO},\ell}) = \frac{1}{B}\sum_{j=1}^{B}\big(\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell} - \bar{\theta}^{*}_{\mathrm{ELBO},\ell}\big)^2, \qquad \bar{\theta}^{*}_{\mathrm{ELBO},\ell} = \frac{1}{B}\sum_{j=1}^{B}\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell}.$$

FIG. 2. *Bootstrap variance estimator.*

CONFIDENCE INTERVAL BY THE BOOTSTRAP PERCENTILE APPROACH.

1. Find the variational estimator $\hat{\theta}_{\mathrm{ELBO}}$ using $X_1, \ldots, X_n$.
2. For $j = 1, \ldots, B$, do the following task:

(a) Sample with replacement from $X_1, \ldots, X_n$ to obtain $X_1^{*(j)}, \ldots, X_n^{*(j)}$.
(b) Compute the variational estimator $\hat{\theta}_{\mathrm{ELBO}}^{*(j)}$ using $X_1^{*(j)}, \ldots, X_n^{*(j)}$.

3. For each parameter, say θ_ℓ, compute $\hat{s}_{\ell,\alpha/2}$ and $\hat{s}_{\ell,1-\alpha/2}$ from

$$\hat{s}_{\ell,\gamma} = \hat{G}_\ell^{-1}(1-\gamma), \qquad \hat{G}_\ell(t) = \frac{1}{B}\sum_{j=1}^{B} I\big(\hat{\theta}_{\mathrm{ELBO},\ell}^{*(j)} - \hat{\theta}_{\mathrm{ELBO},\ell} \le t\big).$$

4. Form the confidence interval:

$$C_{n,\alpha,\ell} = [\hat{\theta}_{\mathrm{ELBO},\ell} + \hat{s}_{\ell,\alpha/2}, \hat{\theta}_{\mathrm{ELBO},\ell} + \hat{s}_{\ell,1-\alpha/2}].$$

FIG. 3. *Confidence interval by the bootstrap percentile approach.*

Equation (8) presents a CI that uses the percentile of the bootstrap distribution of the ELBO estimator. This CI is based on the following approximation:

$$P\big(\hat{\theta}_{\mathrm{ELBO},\ell}^{*} - \hat{\theta}_{\mathrm{ELBO},\ell} < t | X_1, \ldots, X_n\big) \approx P\big(\hat{\theta}_{\mathrm{ELBO},\ell} - \theta_{\mathrm{ELBO},\ell} < t\big). \tag{9}$$

Namely, the CDF of the difference between ELBO estimator and the truth $\hat{\theta}_{\mathrm{ELBO},\ell} - \theta_{\mathrm{ELBO},\ell}$ can be approximated by the CDF of the bootstrapped differences. Thus, $\hat{G}_\ell$ approximates the distribution of the actual difference and we use it to construct a $(1-\alpha)$ CI. We will show the validity of equation (9) in Theorem 2 and Theorem 3.

3.3. *Confidence interval*: *The pivotal approach*. The (studentized) pivotal approach [Wasserman (2006)], also called a percentile-t approach [Hall (1992)], is another popular method for constructing a CI and may lead to a CI with a higher-order correctness [Hall (1992)].

The pivotal approach requires a consistent estimator of the variance of $\hat{\theta}_{\mathrm{ELBO},\ell}$. Let $\hat{\sigma}^2_{\mathrm{ELBO},\ell}$ be such a consistent estimator. Note that $\hat{\sigma}^2_{\mathrm{ELBO},\ell}$ can be constructed using a sandwich estimator as is described in Hall et al. (2011b) and Westling and McCormick (2015). Then the statistic

$$T_n = \frac{\hat{\theta}_{\mathrm{ELBO},\ell} - \theta_{\mathrm{ELBO},\ell}}{\hat{\sigma}_{\mathrm{ELBO},\ell}}$$

acts as a t-statistic and converges to a standard normal distribution [see, e.g., equation (4)]. Therefore, T_n is a pivotal quantity that has asymptotic normality and the pivotal approach is based on bootstrapping T_n to construct a CI.

Instead of using the percentile from a standard normal distribution, we use the bootstrap percentile of T_n. For the jth bootstrap sample, we not only compute

the bootstrap parameter estimate $\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell}$ but also re-compute the corresponding variance estimator $\hat{\sigma}^{*2(j)}_{\mathrm{ELBO},\ell}$ to evaluate the bootstrap version of the pivotal statistics

$$T_n^{*(j)} = \frac{\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell} - \hat{\theta}_{\mathrm{ELBO},\ell}}{\hat{\sigma}^{*(j)}_{\mathrm{ELBO},\ell}}, \qquad j = 1, \ldots, B.$$

We then pick the value $\hat{t}_{\ell,1-\alpha/2}$ as the $(1-\alpha/2)$ upper quantile of the empirical distribution function of $|T_n^{*(1)}|, \ldots, |T_n^{*(B)}|$, that is,

$$\hat{t}_{\ell,1-\alpha/2} = \hat{F}_\ell^{-1}(1-\alpha/2), \qquad \hat{F}_\ell(t) = \frac{1}{B}\sum_{j=1}^{B} I\big(|T_n^{*(j)}| \le t\big).$$

The $(1-\alpha)$ CI is

$$C^{\dagger}_{n,\alpha,\ell} = [\hat{\theta}_{\mathrm{ELBO},\ell} - \hat{\sigma}_{\mathrm{ELBO},\ell} \cdot \hat{t}_{\ell,1-\alpha/2}, \hat{\theta}_{\mathrm{ELBO},\ell} + \hat{\sigma}_{\mathrm{ELBO},\ell} \cdot \hat{t}_{\ell,1-\alpha/2}]. \tag{10}$$

Note that $\hat{\sigma}_{\mathrm{ELBO},\ell}$ is the estimator of the variance of $\hat{\theta}_{\mathrm{ELBO},\ell}$ using the original sample. Figure 4 provides a summary of the bootstrap pivotal approach for constructing a CI.

The intuition of the bootstrap studentized pivotal approach is that the distribution of bootstrap statistic T_n^* (given $X_1, \ldots, X_n$) converges to the distribution of T_n faster than the convergence of T_n to a standard normal distribution. Thus, the CI in equation (10) has a higher order correctness [Babu and Singh (1983), Hall (1992), Singh (1981)].

REMARK 2 (Parametric bootstrap). In addition to the above bootstrap methods, the parametric bootstrap is another popular approach which generates bootstrap samples from $P_{\hat{\theta}_{\mathrm{ELBO}}}$ instead of the empirical distribution function. However, we caution against using the parametric bootstrap. When using the variational estimator, the parametric bootstrap may not give a CI with the (asymptotic) nominal coverage even if the parametric family is correct (i.e., there exists $\theta_0 \in \Theta$ such that the data generating distribution $P = P_{\theta_0}$) because the ELBO estimator $\hat{\theta}_{\mathrm{ELBO}}$ does not converge to θ_0 in general. Thus, $P_{\hat{\theta}_{\mathrm{ELBO}}}$ will not be close to P_{θ_0}, so there is no guarantee that the CI will have nominal coverage. However, when the model is correctly specified and $\hat{\theta}_{\mathrm{ELBO}}$ does converge to θ_0 (this may occur when we allow s to increase; see Remark 5), the parametric bootstrap can provide CIs with nominal coverage; see Bickel et al. (2013) for an example in the case of stochastic block model.

REMARK 3 (Label switching problem). In some models, the MLE may only be unique up to permutation of indices (see, e.g., the example in Section 5). In this case, the ELBO is non-convex so we need to use a gradient ascent method such as the EM algorithm. For each bootstrap sample, we will apply the EM algorithm

CONFIDENCE INTERVAL BY THE BOOTSTRAP PIVOTAL APPROACH.

1. Find the variational estimator $\hat{\theta}_{\mathrm{ELBO}}$ using $X_1, \ldots, X_n$.
2. Compute the variance estimator $\hat{\sigma}^2_{\mathrm{ELBO}} = (\hat{\sigma}^2_{\mathrm{ELBO},1}, \ldots, \hat{\sigma}^2_{\mathrm{ELBO},p})$ using a sandwich estimator.
3. For $j = 1, \ldots, B$, do the following task:

(a) Sample with replacement from $X_1, \ldots, X_n$ to obtain $X_1^{*(j)}, \ldots, X_n^{*(j)}$.
(b) Compute the variational estimator $\hat{\theta}^{*(j)}_{\mathrm{ELBO}}$ using $X_1^{*(j)}, \ldots, X_n^{*(j)}$.
(c) Compute the variance estimator $\hat{\sigma}^{*2(j)}_{\mathrm{ELBO}}$ using $X_1^{*(j)}, \ldots, X_n^{*(j)}$.

4. For each parameter, say θ_ℓ, compute $T_n^{*(1)}, \ldots, T_n^{*(B)}$ using

$$T_n^{*(j)} = \frac{\hat{\theta}^{*(j)}_{\mathrm{ELBO},\ell} - \hat{\theta}_{\mathrm{ELBO},\ell}}{\hat{\sigma}^{*(j)}_{\mathrm{ELBO},\ell}}.$$

5. Compute $\hat{t}_{\ell,\alpha/2}$ and $\hat{t}_{\ell,1-\alpha/2}$ from

$$\hat{t}_{\ell,1-\gamma} = \hat{F}_\ell^{-1}(1-\gamma), \qquad \hat{F}_\ell(t) = \frac{1}{B}\sum_{j=1}^{B} I\big(|T_n^{*(j)}| \le t\big).$$

6. Form the confidence interval:

$$C^{\dagger}_{n,\alpha,\ell} = [\hat{\theta}_{\mathrm{ELBO},\ell} - \hat{\sigma}_{\mathrm{ELBO},\ell} \cdot \hat{t}_{\ell,1-\alpha/2}, \qquad \hat{\theta}_{\mathrm{ELBO},\ell} + \hat{\sigma}_{\mathrm{ELBO},\ell} \cdot \hat{t}_{\ell,1-\alpha/2}].$$

FIG. 4. *Confidence interval by the bootstrap pivotal approach.*

with the same initialization (we recommend to use the estimator of the original sample as the initial point for each bootstrap sample). This will avoid the problem of label switching [Redner and Walker (1984)] and the bootstrap will recover the uncertainty in parameter estimation.

4. Asymptotic distribution and bootstrap consistency. In this section, we derive the asymptotic distribution of the variational estimator and its bootstrap theory. We will study the theory in both scenarios: fixing and increasing d, the dimension of parameters θ, after introducing further notation.

Let $B(x, r)$ be a ball with radius r centered at x. We define $\Psi(\theta) = \mathbb{E}(\mathcal{E}(\theta|X_1))$, and let $\Psi_\theta = \nabla\Psi$ and $\Psi_{\theta\theta} = \nabla\nabla\Psi$ to be the gradient and Hessian matrix of Ψ, respectively. For a unit vector $b \in \mathbb{R}^d$ and a function $f : \mathbb{R}^d \mapsto \mathbb{R}$, $\nabla_b f = b^T \nabla f$ is the derivative of f in the direction of b. For a matrix $A \in \mathbb{R}^{d\times d}$, we denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ to be the largest and the smallest eigenvalues of A, respectively.

4.1. *Fixed dimension.* When the dimension d is fixed, the ELBO estimator and its target can be analyzed using the theory of M-estimators [van der Vaart

(1998)]. The asymptotic normality of $\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}}$ has been analyzed in the literature [Bickel et al. (2013), Hall et al. (2011b), Wang and Blei (2017), Westling and McCormick (2015)] under several scenarios. Here we present the asymptotic normality using the result stated in Westling and McCormick (2015) because they also considered frequentist estimation in the general context of latent variable models.

THEOREM 1 (Theorem 2 in Westling and McCormick (2015)). *Assume conditions* (B1)–(B5) *in the Appendix of Westling and McCormick* (2015). *Then*

$$\sqrt{n}(\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}}) \xrightarrow{D} N\big(0, V(P_0, \theta_{\mathrm{ELBO}})\big), \tag{11}$$

where $V(P_0, \theta_{\mathrm{ELBO}}) = A(P_0, \theta_{\mathrm{ELBO}})^{-1} B(P_0, \theta_{\mathrm{ELBO}}) A(P_0, \theta_{\mathrm{ELBO}})$ *is a* $p \times p$ *matrix such that*

$$A(P_0, \theta_{\mathrm{ELBO}}) = \mathbb{E}_{X \sim P_0}\big(\Psi_{\theta\theta}(\theta_{\mathrm{ELBO}}|X)\big),$$

$$B(P_0, \theta_{\mathrm{ELBO}}) = \mathbb{E}_{X \sim P_0}\big(\Psi_{\theta}(\theta_{\mathrm{ELBO}}|X)\Psi_{\theta}(\theta_{\mathrm{ELBO}}|X)^T\big).$$

We include the assumptions (B1)–(B5) in the Appendix of Westling and McCormick (2015) in Appendix B. These assumptions are made to derive the asymptotic normality of an M-estimator [see, e.g., Theorem 5.23 of van der Vaart (1998)]. Essentially, these assumptions assure that $\omega_{\max}(\theta|x)$, $\mathsf{ELBO}(\theta, \omega|x)$, and $\mathcal{E}(\theta|x)$ are well-defined and sufficiently smooth and well-behaved around θ_{ELBO} and P_0-a.e. x. Viewing the ELBO estimator as the MLE, the quantity $\Psi_\theta(\cdot)$ and $V(P, \theta_{\mathrm{ELBO}})$ are analogous to the score function and the Fisher information matrix, respectively.

To describe the validity of a bootstrap procedure, we often use the notion of convergence under Kolmogorov distance [van der Vaart (1998)]. For two random variables A and B, their Kolmogorov distance is

$$\sup_t \big|P(A < t) - P(B < t)\big|.$$

The bound on Kolmogorov distance is also called the Berry–Esseen bound [Berry (1941), Esseen (1942)]. Note that convergence in probability in Kolmogorov distance is a stronger result, compared to convergence in distribution. Namely, if a sequence of random variables $A_1, \ldots, A_n, \ldots$ with $d_K(A_n, A_0) \xrightarrow{P} 0$, then $A_n \xrightarrow{D} A_0$.

Let $\Delta_n = \sqrt{n}(\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}})$ and $\Delta_n^* = \sqrt{n}(\hat{\theta}^*_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})$ be the scaled difference and the bootstrap version of it. We will prove that Δ_n and Δ_n^* converge in Kolmogorov distance.

THEOREM 2. *Assume conditions* (B1)–(B5) *in the Appendix of Westling and McCormick* (2015) *and* $\mathbb{E}\|\Psi_\theta(\theta_{\mathrm{ELBO}}|X_1)\|^3 < \infty$. *Then for any vector* $a \in \mathbb{R}^d$ *such that* $\|a\| = 1$,

$$\sup_t \big|P\big(a^T \Delta_n^* < t | X_1, \ldots, X_n\big) - P\big(a^T \Delta_n < t\big)\big| \xrightarrow{P} 0. \tag{12}$$

Thus, for any $\ell = 1, \ldots, d$,

$$P(\theta_\ell \in C_{n,\alpha,\ell}) \to 1 - \alpha,$$

$$P\big(\theta_\ell \in C^{\dagger}_{n,\alpha,\ell}\big) \to 1 - \alpha,$$

where $C_{n,\alpha,\ell}$ *and* $C^{\dagger}_{n,\alpha,\ell}$ *are the CIs based on equations* (8) *and* (10), *respectively.*

The proof is deferred to the Appendix. Theorem 2 shows that no matter which orientation we project onto (using the unit vector a), the distribution of random variable $\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}}$ and the distribution of its bootstrap variant $\hat{\theta}^*_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}$ converge. Thus, the bootstrap quantile converges to the quantile of the actual distribution, which proves validity of the bootstrap.

4.2. *Increasing dimension*. We now study the bootstrap theory when the dimension of parameters is allowed to increase with respect to the sample size. These situations occurs in many scenarios. For example, in a mixed membership model, we may want to increase the number of subgroups when we have a larger sample. Or in an item response theory model, both the number of questions in a test and the number of participants may be increasing at the same time [Douglas (1997), Haberman (1977)]. In this case, we will write $d = d_n \to \infty$ as $n \to \infty$. Note that we only allow d, the dimension of θ increase and the dimension of variational parameters are assumed to be fixed. Thus, the population quantity θ_{ELBO} will also be changing.

Assumptions.

(A0) $\theta_{\mathrm{ELBO}} \in \Theta$ is the unique maximizer of $\Psi(\theta)$ and $\omega_{\max}(\theta|x)$ is unique for each $\theta \in \Theta$ and almost surely for $x \in \mathbb{R}^J$ under P_0.

(A1) There exists $c_0 > 0$ such that all eigenvalues of $\Psi_{\theta\theta}(\theta_{\mathrm{ELBO}})$ are not in $[-c_0, c_0]$ for any d.

(A2) There exists $r_0, c_1 > 0$ such that for any unit vectors $b_1, b_2, b_3 \in \mathbb{R}^d$,

$$\sup_x \big|\nabla_{b_1}\nabla_{b_2}\nabla_{b_3}\mathcal{E}(\theta|x)\big| \le c_1 < \infty$$

for all $\theta \in B(\theta_{\mathrm{ELBO}}, r_0)$ and d.

(A3) There exists $c_2 > 0$ such that for any unit vector $a \in \mathbb{R}^d$,

$$\mathbb{E}\big(\big|\nabla_a \mathcal{E}(\theta|X_1)\big|^3\big) \le c_2 < \infty$$

for any d.

(A0) is a very common assumption that requires θ_{ELBO} to be uniquely defined [Westling and McCormick (2015)]. Note that we can relax (A0) to require θ_{ELBO} to be unique under permuting the indices when the model is symmetric (such as the example in Section 5). The theoretical results will be the same after a small

modification to the proof so here we make this assumption to simplify the exposition.

(A1) implies that the Hessian matrix of Ψ is invertible at θ_{ELBO} when $n, d \to \infty$. This is a generalization of the invertible Fisher information matrix condition to the increasing-dimensional setting.

(A2) can be viewed as a generalization of a bounded 2-norm of the third derivative tensor $\nabla\nabla\nabla\mathcal{E}(\theta_{\mathrm{ELBO}}|x)$. To see this, consider only two-directional derivative, $|\nabla_{b_1}\nabla_{b_2}\mathcal{E}(\theta_{\mathrm{ELBO}}|x)|$. The supremum of this will be the 2-norm (maximum absolute eigenvalue) of $\mathcal{E}(\theta_{\mathrm{ELBO}}|x)$. Note that assumptions similar to (A1–2) also appear in Portnoy (1985) and Mammen (1989).

(A3) is a third moment condition that is used to establish a Berry–Esseen bound [Berry (1941), Esseen (1942)]. Note that when d is changing with respect to n, (A2) and (A3) can be relaxed so that constants c_1 and c_2 can depend on n. However, this relaxation will put another constraint on how fast $d \to n$ with respect to $n \to \infty$.

Note that we do not assume the distribution $P_\theta = P(x, z; \theta)$ belongs to an exponential family. If P_θ belongs the exponential family, the assumptions can be weakened to the assumptions in Portnoy (1988).

THEOREM 3. *Assume* (A0)–(A3) *and* $d = d_n \to \infty$ *and* $\frac{d^2}{n} \to 0$. *Then, for any vector* $a_n \in \mathbb{R}^d$ *such that* $\|a_n\| = 1$, *there exists a number* $v(a_n)$ *such that*

$$\sup_t \left| P\left(a_n^T \Delta_n < t\right) - P\left(\sigma(a_n) \cdot Z < t\right)\right| \to 0, \tag{13}$$

where Z *is a standard normal random variable. Moreover,*

$$\sup_t \left| P\left(a_n^T \Delta_n^* < t | \mathcal{X}\right) - P\left(a_n^T \Delta_n < t\right)\right| \xrightarrow{P} 0. \tag{14}$$

Thus, for any $\ell = 1, \ldots, d$,

$$P(\theta_\ell \in C_{n,\alpha,\ell}) \to 1 - \alpha$$

$$P\left(\theta_\ell \in C^\dagger_{n,\alpha,\ell}\right) \to 1 - \alpha,$$

where $C_{n,\alpha,\ell}$ *and* $C^\dagger_{n,\alpha,\ell}$ *are the CIs based on equations* (8) *and* (10), *respectively.*

The proof is deferred to the Appendix. The first assertion in Theorem 3 states that the difference between the ELBO estimator and its target converges to a normal distribution when we project the difference to any direction. The quantity $\sigma(a)$ is the standard deviation of the difference of the estimator that depends on the data-generating distribution P_0 and on the variational family that is being used. Note that, when the dimension is fixed, $\sigma(a) = a^T V(P_0, \theta_{\mathrm{ELBO}})a$.

The second assertion in Theorem 3 shows that the limiting distributions of the scaled difference and its bootstrap variant are asymptotically the same. This implies that the CI constructed using the bootstrap or variance estimated by the bootstrap is asymptotically valid.

Note that the requirement $\frac{d^2}{n} \to 0$ is very common in increasing-dimensional problem; see, for example, Mammen (1993), Portnoy (1988).

REMARK 4 (Increasing both d and s). Theorem 3 can be applied to a case where both the dimension of parameter d and the dimension of variational parameter s are increasing. In this case, we need assumptions (A0)–(A3) to hold for every s and d. When we allow $s = s_n$ to increase, the assumption (A1) may be too strong. We can relax this assumption by allowing the constant c_0 in (A1) to decrease to 0 slowly. The increasing rate of s_n will be constrained by the decreasing rate of c_0 to guarantee the invertibility of $\Psi_{\theta\theta}$.

REMARK 5 (Increasing s only). Even when d, the dimension of the parameter, remains fixed (i.e., the population MLE θ_{MLE} is fixed), changing s, the dimension of ω, will also change the (population) quantity $\theta_{\mathrm{ELBO}} = \theta_{\mathrm{ELBO},s}$. In some situations, we even have $\theta_{\mathrm{ELBO}} = \theta_{\mathrm{ELBO},s} \to \theta_{\mathrm{MLE}}$; see Hall et al. (2011b) and Bickel et al. (2013) for examples. The difference $\theta_{\mathrm{ELBO},s} - \theta_{\mathrm{MLE}}$ can be viewed as the bias of the variational estimator. Because the dimension of variational parameter s can be viewed as a measure of model complexity of the variational estimator, the property $\theta_{\mathrm{ELBO},s} \to \theta_{\mathrm{MLE}}$ can be interpreted as an asymptotic unbiasedness property in terms of model complexity.

REMARK 6 (High-dimensional case). When $d > n$, the conventional central limiting theorem fails because of the complexity coming from the high dimensional parameters [Portnoy (1984, 1985)]. Thus, CIs from the percentile or pivotal approaches do not have the nominal coverage. However, it is still possible to construct an asymptotically valid CI using the bootstrap. The rectangle CI [Chernozhukov, Chetverikov and Kato (2013)] is one example. We refer the readers to Chernozhukov, Chetverikov and Kato (2013), Fan and Zhou (2016), Wasserman, Kolar and Rinaldo (2013) for more details about rectangle CIs.

5. Data analysis. We illustrate our theoretical results with multivariate binary data on functional disability from the National Long Term Care Survey (NLTCS). Erosheva, Fienberg and Joutard (2007) presented the first case of variational estimation for mixed membership models with binary data from the NLTCS. Here, we consider observations collected on the NLTCS participants in 1984, 1989, and 1994. The data contain binary indicators on six activities of daily living (ADL) and 10 instrumental activities of daily living (IADL) for community-dwelling elderly. The six ADL items include basic activities of hygiene and personal care: eating, getting in/out of bed, getting around inside, dressing, bathing, and getting to the bathroom or using toilet. The 10 IADL items include basic activities necessary to reside in the community: doing heavy housework, doing light housework, doing laundry, cooking, grocery shopping, getting about outside, traveling, managing money, taking medicine, and telephoning. Responses are coded as 0 and 1,

where 1 denotes a presence and 0 denotes an absence of a functional disability. In the NLTCS, positive (1) ADL responses mean that during the past week the activity had not been, or was not expected to be, performed without the aid of another person or the use of equipment; negative (0) IADL responses mean that a person usually could not, or was not going to be able to, perform the activity because of a disability or a health problem. For a more in-depth discussion, see Manton, Corder and Stallard (1993) and Erosheva and White (2006).

Similar to Erosheva, Fienberg and Joutard (2007), we also use a mixed membership analysis. Erosheva, Fienberg and Joutard (2007) take a fully Bayesian approach and specify priors for the α and Π model parameters discussed below; however, in this analysis we take a frequentist approach and directly compute maximum ELBO estimates for α and Π. Also, Erosheva, Fienberg and Joutard (2007) analyze all four waves (1982, 1984, 1989, and 1994), but we restrict our analysis to the 1984, 1989, and 1994 waves.

In particular, we are interested in two tasks. First, we use a mixed membership model to describe the 5934 observations in the 1984 wave. We use a variational procedure to estimate the model parameters and then give bootstrapped confidence intervals for each of those estimates. Next, we consider the 4463 and 5089 observations from 1989 and 1994 respectively. We test whether the responses observed in 1989 and 1994 arise from the same distribution. Given the two natural sub-populations, this corresponds to a possible two-sample test described in Section 2.1. A conceptually simpler approach could be used instead of a model based approach. At the coarsest resolution, this might be a two sample t-test for each of the 16 variables, and at the finest resolution, this might be a two sample t-test for each of the 2^{16} possible response patterns. However, testing in the mixed membership framework allows investigation of subtle changes in the underlying structure, while still retaining easy interpretation.

5.1. *Mixed membership models and variational inference.* Throughout this analysis, we use mixed membership models to uncover latent structure. Like a mixture model, mixed membership models assume that the population is comprised of several groups, where each group has a distribution over the observed variables. However, while mixture models assume that each individual belongs to a single group, mixed membership models allow each individual to have a partial membership in multiple groups [Airoldi et al. (2015)]. Mixed membership models have been used for topic modeling [Blei, Ng and Jordan (2003)], social network analysis [Airoldi et al. (2008)], survey data [Erosheva, Fienberg and Joutard (2007)], and statistical genetics [Pritchard, Stephens and Donnelly (2000)]. Note that allowing for mixed membership differs from estimating the posterior probability of group assignment when using a mixture model. Under a mixture model, as the data about an individual grows, the posterior should concentrate on a single group, while in a mixed membership model, as the data about an individual grows, we may consistently estimate the individual's membership, which could be in the interior of the simplex.

In the setting we consider, for each individual $i = 1, \ldots, n$ we observe multivariate data $X_i = (X_{i,1}, \ldots, X_{i,16})$ and assume the following generative model. Let $j = 1, \ldots, J = 16$ index variables and K be the fixed number of groups. We assume fixed parameters $\alpha \in \mathbb{R}^K_{>0}$, which regulates the Dirichlet distribution for group membership, and $\Pi = \{\pi_{jk}\}$ for $j = 1, \ldots, J$ and $k = 1, \ldots, K$, where π_{jk} is the Bernoulli parameter for a response to variable j from a full member of group k. The generative model for individual i is:

1. $\lambda_i \sim \text{Dirichlet}(\alpha)$, where λ_i lies in the $K-1$ simplex (i.e., $\sum_k \lambda_{ik} = 1$ and $\lambda_{ik} \geq 0$). Each element λ_{ik} characterizes the extent of membership for individual i in group k.
2. For each variable j:

- $g_{ij} \sim \text{Categorical}(\lambda_i)$, where $g_{ij} \in \{1, \ldots, K\}$ indicates the group whose parameters govern individual i's response to question j.
- $X_{ij} \sim \text{Bernoulli}(\pi_{jg_{ij}})$, the observed response for individual i on question j.

This hierarchical model assumes that each individual responds to each question as a full member of group g_{ij}. However, for each individual, the group may vary across variables and the probability of responding as a full member of group k for each question is governed by λ_{ik}. In addition, X_{ij} is independent of $X_{ij'}$ given λ_{ik}.

The parameters of interest are α and Π. For the Dirichlet parameter α, the quantity $\alpha_k / \sum_{k'} \alpha_{k'}$ indicates the relative proportion of each group and the magnitude, $\sum_k \alpha_k$, indicates the level of intra-individual mixing. Distributions with larger values of $\sum_k \alpha_k$ concentrate density in the interior of the simplex and imply a higher level of intra-individual mixing, while distributions with smaller values of $\sum_k \alpha_k$ concentrate density in the corners of the simplex and indicate less intra-individual mixing. The Bernoulli parameters Π characterize the ability/disability of each group. The parameters λ_i and g_{ij} are latent variables which we consider as nuisance parameters. In the previous notation, $\theta = \{\alpha, \Pi\}$ and $Z_i = \{\lambda_i, g_{ij}\}$.

Although the model is straightforward to describe and generate, maximum likelihood estimation is difficult because the normalizing constant is intractable. Thus, to fit the model, we use the `mixedMem` R package [Wang and Erosheva (2015)] which specifies the following mean field variational distribution with variational parameters $\phi_i \in \mathbb{R}^K_{>0}$ and δ_{ij} in the $K-1$ simplex:

$$\begin{aligned} \lambda_i &\sim \text{Dirichlet}(\phi_i); \\ g_{ij} &\sim \text{Categorical}(\delta_{ij}); \\ X_{ij} &\sim \text{Bernoulli}(\pi_{jg_{ij}}). \end{aligned} \tag{15}$$

In the previous notation, $\omega_i = \{\phi_i, \delta_{ij}\}$. The likelihood and specified variational

distribution yield the following ELBO:

$$
\begin{aligned}
&\mathsf{ELBO}(\theta, \omega | X) \\
&\quad = \sum_i \log\Gamma\Big(\sum_k \alpha_k\Big) - \sum_{i,k} \log\Gamma(\alpha_k) \\
&\qquad + \sum_{i,k} (\alpha_k - 1)\Big[\Psi(\phi_{ik}) - \Psi\Big(\sum_k \phi_{ik}\Big)\Big] \\
&\qquad + \sum_{i,j,k} \delta_{ijk}\Big[\Psi(\phi_{ik}) - \Psi\Big(\sum_k \phi_{ik}\Big)\Big] \\
&\qquad + \sum_{i,j,k} \delta_{ijk} X_{ij} \log(\pi_{jk}) + \sum_{i,j,k} \delta_{ijk}(1 - X_{i,j}) \log(1 - \pi_{jkv}) \\
&\qquad - \sum_i \log\Gamma\Big(\sum_k \phi_{ik}\Big) + \sum_{i,k} \log\Gamma(\phi_{ik}) \\
&\qquad - \sum_{i,k} (\phi_{ik} - 1)\Big[\Psi(\phi_{ik}) - \Psi\Big(\sum_k \phi_{ik}\Big)\Big] - \sum_{i,j,k} \delta_{ijk} \log(\delta_{ijk}),
\end{aligned} \tag{16}
$$

where $\Gamma(\cdot)$ is the gamma function and $\Psi(\cdot)$ is the digamma function which is the derivative of the log-Γ function. We maximize the ELBO with respect to the parameters of interest, α and Π, and the variational parameters, ϕ_i and δ_{ij}, through a block coordinate ascent procedure which alternates between two steps. In the first step, holding α and Π fixed, we compute the optimal variational parameters by iterative coordinate ascent. Then, holding the variational parameters fixed, we update α and Π through a Newton–Raphson procedure. Because there is no closed-form solution for $\hat{\delta}_{ij}(\alpha, \Pi)$ and $\hat{\phi}_i(\alpha, \Pi)$, we can not easily compute a Hessian required for the sandwich estimator of Westling and McCormick (2015) or the pivotal confidence intervals summarized by Figure 4. However, percentile based bootstrap confidence intervals and bootstrap variance estimates can be used.

5.2. *Initial analysis and bootstrapped standard errors*. We first select an appropriate number of groups, K, using a pseudo-BIC criterion:

$$
\text{pBIC} = p \log(n) - 2 \times \mathsf{ELBO}(\hat{\theta}_{\text{ELBO}}, \hat{\omega}_{\text{ELBO}} | X), \tag{17}
$$

where $p = K + J \times K$, the count of parameters α and Π. Because the ELBO is generally multi-modal, we use 1000 random initialization points (for α and Π) for $K = 2, \ldots, 9$. For each K, we then select the resulting stationary point with the largest ELBO and compute the pseudo-BIC. Using many random restarts is important, because, as is typically the case, the ELBO defined by the mixed membership model and variational family we use is multi-modal. We see from the left panel of Figure 5 that the ELBO of each stationary point can vary widely within

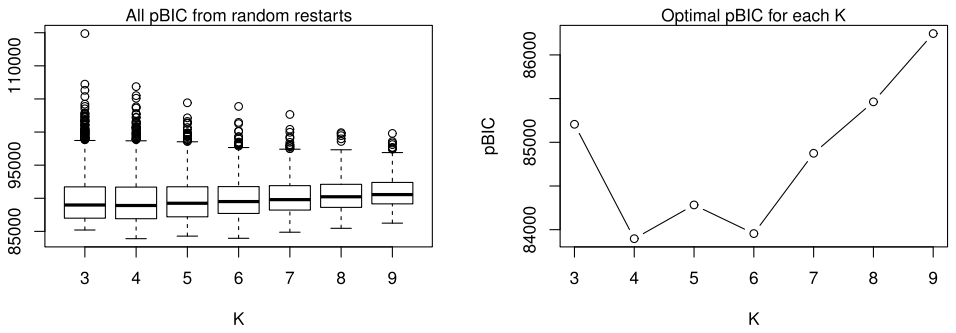

FIG. 5. *The pseudo-BIC across levels of K groups. The left panel shows pseudo-BIC values across all random initializations, and the right panel shows only the optimal model for each K.*

each value of K. In the right panel of Figure 5, we plot only the lowest pBIC for each K; we see that the pBIC criteria leads us to select a 4 group model, though a 6 group model might also be appropriate.

The estimated Bernoulli and Dirchlet parameters for the optimal 4 group model are presented in Figure 6 and 7. The confidence intervals shown in black are calculated by using the nonparametric percentile bootstrap summarized in Figure 3 and the confidence intervals shown in red are calculated using the parametric percentile bootstrap. The intervals used are post model-selection [Leeb and Pötscher (2005)].

Since the ELBO can be multi-modal and we use a coordinate ascent procedure, we need to carefully initialize each bootstrap run so that we do not enter another basin of attraction and overestimate the sampling variability. In particular, we initialize the global parameters, α and Π, as well as the individual latent variables, λ_i and g_{ij}, at the corresponding quantities estimated from the original data. In general, we expect each bootstrap run to require less computational effort than the original estimation procedure since we expect the initialization to be near the stationary point. In addition, each of the bootstrap runs can be easily parallelized on a cluster; for this particular analysis, an individual bootstrap run took roughly 15 seconds on a laptop.

In Figure 6, we have sorted the groups top to bottom (1 through 4) by least disabled to most disabled. Group 1 is generally most likely to be able to perform each of the 16 tasks. Group 2 appears to be relatively less able to perform most physical/mobility related tasks, but is relatively more able to perform tasks requiring mental acuity. For instance, members of Group 2 are relatively less able to get in/out of bed, move around inside, and move around outside; however, they are relatively more able to cook, manage money, take medicine, and use the telephone. Group 3 appears to have more mobility, but is less able to perform tasks which require mental acuity. For instance, individuals in Group 3 are relatively more able to get in/out of bed, move around inside, and use the toilet, but less able to manage

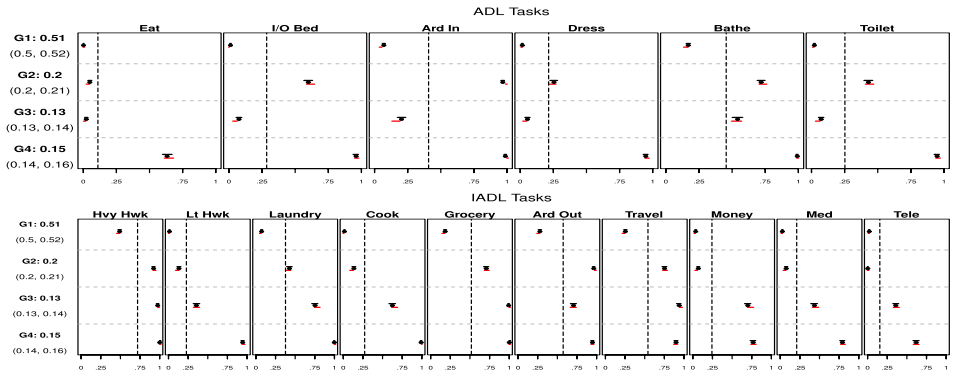

FIG. 6. *The estimated parameters and CIs for the* 4 *group mixed membership model. The black CI's are formed using the nonparametric bootstrap; the red CI's are formed using the parametric bootstrap. The top panel shows estimates for the ADL activities and the bottom panel shows estimates for the IADL activities. The estimated population proportion,* $\hat{\alpha}_k / \sum_{k'} \hat{\alpha}_{k'}$ *is shown on the left with the corresponding CI under each group label. For aiding interpretation, the vertical dashed lines shows the marginal proportion of individuals whose response was* 1 *for each variable.*

money or use the telephone. Finally Group 4 is generally least likely to be able to perform each task. The estimated Bernoulli parameters for Group 4 are higher than the marginal probabilities for all 16 tasks. Note that the CIs from the two bootstrap methods are small, indicating that our estimators are quite precise.

We caution against using the parametric bootstrap with variational inference since θ_{ELBO} in general is not equal to θ_{MLE} so the CI's may not always cover the variational point estimates. In particular, for the Bernoulli parameters, 37 of the 64 CI's constructed via the parametric bootstrap do not cover the point estimates. In addition, all four of the parametric bootstrap CI's for $\hat{\alpha}$ (and 2/4 of the CI's for the

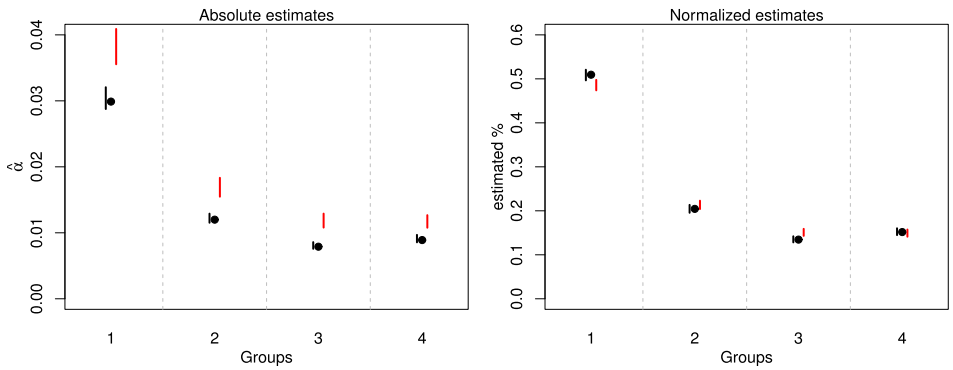

FIG. 7. *The left panel shows the estimates and CI's for* α *and the right panel shows the estimates and CI's for the proportion of each group; that is,* $\alpha_k / \sum_{k'} \alpha_{k'}$. *The black CI's are formed using the nonparametric bootstrap and the red CI's are formed using the parametric bootstrap.*

population proportions) do not cover the point estimates. However, all of the CI's (both for the Bernoulli and Dirichlet parameters) constructed using the nonparametric bootstrap do cover the point estimates. As noted by Andrews (2000), when the point estimates are near the boundary of the parameter space, the bootstrap estimates might be unstable. In this case, we see that when the Bernoulli parameters are close to 0 or 1, this generally causes a problem for the parametric bootstrap, but not for the nonparametric bootstrap.

5.3. *Two-sample test.* We now consider observations from the 1989 and 1994 waves. In particular, we test whether the functional disability measures taken five years apart are generated by the same distribution. In order to concretely interpret differences between the two waves, we fix $K = 4$ and use the Bernoulli parameters estimated from the 1984 wave. We then find point estimates $\hat{\alpha}_{1989}$ and $\hat{\alpha}_{1994}$ separately by maximizing the ELBO with respect to α (keeping Π fixed). Again, because of multi-modailty of the ELBO, we use 1000 random initialization to select an $\hat{\alpha}$ for each wave. In principal, fixing the Bernoulli parameters to any random quantity and concluding that $\alpha_{89,\text{ELBO}|\Pi} \neq \alpha_{94,\text{ELBO}|\Pi}$ would result in rejecting the null hypothesis (where $\alpha_{\text{ELBO}|\Pi}$ indicates the α value which maximizes the ELBO for fixed Π). However, we use the point estimates from the 1984 wave to facilitate interpretability.

The estimated group proportions for 1989 and 1994 are shown in Figure 8 with the corresponding confidence intervals formed by the nonparametric percentile bootstrap standard errors. In 1994, the prevalence of the least disabled group (Group 1) increased, while the prevalence of Group 2 (incapable of mobility tasks, but capable of mental tasks), Group 3 (capable of mobility tasks, but incapable of mental tasks), and Group 4 (generally incapable of all tasks) all decreased by roughly 0.03 each.

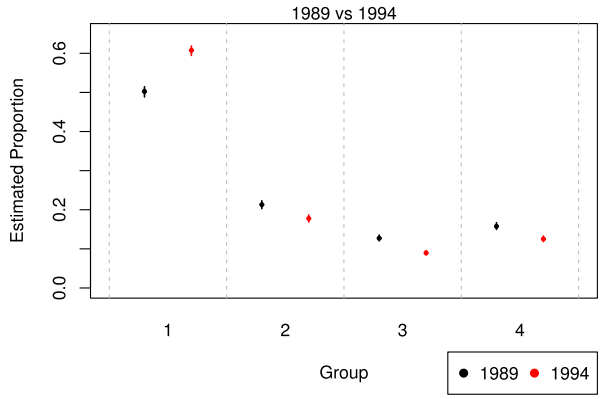

FIG. 8. *The estimated proportions of each group for* 1989 *and* 1994. *The CIs are percentile method bootstrapped intervals.*

For good measure, we also perform a Wald test for the population proportions:

$$\hat{p} = (\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3) \Big/ \sum_{k=1}^{4} \hat{\alpha}_k,$$

where the proportion for Group 4 is excluded so that the distribution is non-degenerate. Using the bootstrap estimate of covariance $\widehat{V(\hat{p})}$ calculated by the procedure summarized in Figure 2, we find that

$$(\hat{p}_{1989} - \hat{p}_{1994})^T \big(\widehat{V(\hat{p}_{1989})} + \widehat{V(\hat{p}_{1994})}\big)^{-1} (\hat{p}_{1989} - \hat{p}_{1994}) = 143.7.$$

Thus, we reject the null hypothesis that all population proportions are equal with a p-value less than 10^{-16} when compared to the χ^2_3 reference distribution.

6. Discussion. We conclude this paper by including some remarks about practical aspects of the variational inference and the bootstrap, and by making several observations on the connections between the research presented in this paper and the work of late Stephen E. Fienberg, to whom this special issue of the *Annals of Applied Statistics* is dedicated.

Bootstrap versus asymptotic normality. For practitioners, constructing a CI using a bootstrap approach is generally easier than using the asymptotic normality [Hall et al. (2011b), Westling and McCormick (2015)]. To construct a CI using asymptotic normality, we need a (consistent) variance estimator and calculating that estimator often requires an involved derivation, which could be very challenging when the model is complex. The NLTCS example of mixed membership is one such example. And sometimes, such an estimator does not exist so we are unable to use the asymptotic normality approach. On the other hand, the implementation of a bootstrap approach is very easy—it is just sampling with replacement and re-applying the variational inference. The bootstrap approach does not require a consistent variation estimator so it is a more general approach than the interval from asymptotic normality approach. Moreover, if we do have a variance estimator, as is discussed in Section 3.3, we can construct a CI using the bootstrap pivotal approach, which may lead to a CI with a higher order correctness [Babu and Singh (1983), Hall (1992), Singh (1981)].

Implications for variational inference in Bayesian settings. Theorems 1 and 2 proved the asymptotic normality and bootstrap validity of using the variational inference to approximate the MLE. These theorems can be applied to Bayesian variational estimators as long as the prior is sufficiently smooth (for a trivial case, consider a uniform prior on the parameter space) or to a penalized ELBO with a very weak (i.e., asymptotically negligible) penalty. However, in the Bayesian framework, the posterior distribution and credible intervals are the quantities of the interest and the CI is not the main objective. For the penalized ELBO, a weak penalty is often not of research interest because it neither encourages sparsity nor stabilizes the estimator.

Two-sample test and comparison. The two-sample test used in Section 5.3 shows great potential of combining the bootstrap CI and variational inference. In the NLTCS example that we presented, without adequate tools to obtain uncertainty in estimates, it is possible that an erroneous conclusion could have been made, stating that the proportion of responses that corresponds to profile 3 (mainly problems with managing money, grocery shopping, and traveling) stays the same over the 10 year interval, while our analysis demonstrated that the difference is significant. Note that the approach of comparing two samples is very generic—it can be applied to various problems involving a comparison of two datasets using variational inference. With the methodologies developed in this paper, we can assess the significance of the difference between estimates using the bootstrap and make a statistical conclusion about the two datasets.

Connection with Stephen E. Fienberg. Stephen Fienberg had originally introduced Erosheva, then a graduate student, to the Grade of Membership Model [Woodbury, Clive and Garson (1978)]and to the functional disability data from the NLTCS. Erosheva's graduate work has motivated the original NLTCS publication [Erosheva, Fienberg and Joutard (2007)] as well as the development of the general mixed membership modeling framework. For that original NLTCS publication, under Fienberg's direction, Erosheva and Joutard have developed and implemented both a fully Bayesian MCMC approach and the corresponding variational estimation algorithm for mixed membership models with binary data [Erosheva, Fienberg and Joutard (2007)]. Later, Wang, Matsueda and Erosheva (2017), extended this line of work and developed a variational estimation algorithm for mixed membership models with rank data, where, at a suggestion of a reviewer, they used bootstrap methods to assess uncertainty in the model estimates.

Although Fienberg's impact on statistical science spanned many areas, including the census and survey research in general, he is perhaps best known for his contributions to discrete data analysis and log-linear models. Even though he used to say that "everything is a log-linear model", meaning that almost any statistical model for discrete data has a log-linear representation, he did not brush off other approaches. In particularly, Stephen Fienberg was a big advocate of mixed membership models—because of their flexibility and practical appeal—during the last 15 years of his career. Mixed membership models present certain challenges in estimation, and, while recommending variational inference as a step toward solving those challenges, Fienberg was very much cognizant of both the practical advantages and the lack of statistical theory for variational estimation. We are not aware of his involvement in recent efforts to provide a theoretical foundation for variational estimates, but we can say with confidence that he would have been supportive and encouraging for advancing research in this direction.

APPENDIX A: PROOFS

PROOF OF THEOREM 2. Our proof consists of two parts. In the first part, we show that the asymptotic normality admits a Berry–Essen bound. In the second part, we show that the bootstrap variant converges to the same distribution with a Berry–Essen bound.

Part 1: Berry–Esseen Bound. By the derivation of Theorem 2 in Westling and McCormick (2015) and Theorem 5.23 in van der Vaart (1998), the ELBO estimator has the property that

$$\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}} = A(P_0, \theta_{\mathrm{ELBO}}) \cdot \frac{1}{n} \sum_{i=1}^{n} \Psi_\theta(\theta_{\mathrm{ELBO}} | X_i) + o_P\left(\frac{1}{\sqrt{n}}\right).$$

Note that the above equation is a common expression for an M-estimator. Thus, $a^T \Delta_n$ has the following expression:

$$\begin{aligned} a^T \Delta_n &= a^T \sqrt{n}(\hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}}) \\ &= a^T A(P_0, \theta_{\mathrm{ELBO}}) \cdot \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \Psi_\theta(\theta_{\mathrm{ELBO}} | X_i) + o_P(1) \\ &= \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \big(W_i - \mathbb{E}(W_i)\big) + o_P(1), \end{aligned}$$

where $W_i = a^T A(P_0, \theta_{\mathrm{ELBO}}) \Psi_\theta(\theta_{\mathrm{ELBO}} | X_i)$ and $\mathbb{E}(W_i) = 0$ because $\mathbb{E}(\Psi_\theta(\theta_{\mathrm{ELBO}} | X_1)) = 0$.

By the assumption that $\mathbb{E}\|\Psi_\theta(\theta_{\mathrm{ELBO}} | X_1)\|^3 < \infty$ and the Berry–Esseen theorem [Berry (1941), Esseen (1942)], we conclude that

$$\sup_t \big| P\big(a^T \Delta_n < t\big) - P(Z_w < t) \big| \le C_{\mathrm{BE}} \frac{\mathbb{E}\|W_1\|^3}{\sqrt{n}} + o_P(1),$$

where Z_w is a normal distribution with variance $\mathsf{Var}(W_1)$ and C_{BE} is a universal constant from the Berry–Esseen theorem.

Part 2: bootstrap. Let $\mathcal{X}_n = \{X_1, \ldots, X_n\}$. For the bootstrap case, we have a similar decomposition of $\hat{\theta}^*_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}$:

$$\hat{\theta}^*_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}} = A(\hat{P}_n, \hat{\theta}_{\mathrm{ELBO}}) \cdot \frac{1}{n} \sum_{i=1}^{n} \Psi_\theta\big(\hat{\theta}_{\mathrm{ELBO}} | X_i^*\big) + E_n^*,$$

where $\|E_n^*\| = o_P(\frac{1}{\sqrt{n}})$ is a small correction error and $\hat{P}_n$ is the empirical distribution. Thus,

$$\begin{aligned} a^T \Delta_n^* &= a^T \sqrt{n}\big(\hat{\theta}^*_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}\big) \\ &= \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \big(W_i^* - \mathbb{E}\big(W_i^* | \mathcal{X}_n\big)\big) + o_P(1), \end{aligned}$$

where $W_i^* = a^T A(\hat{P}_n, \hat{\theta}_{\mathrm{ELBO}})\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_i^*)$ and $\mathbb{E}(W_i^*|\mathcal{X}_n) = 0$. Again, by applying the Berry–Essen theorem [Berry (1941), Esseen (1942)], we conclude that

$$\sup_t \left|P\left(a^T\Delta_n^* < t|\mathcal{X}_n\right) - P\left(Z_w^* < t|\mathcal{X}_n\right)\right| \le C_{\mathrm{BE}}\frac{\hat{\mathbb{E}}_n\|W_1\|^3}{\sqrt{n}} + o_P(1),$$

where Z_w^* is a normal distribution with variance $\mathsf{Var}(W_1^*|\mathcal{X}_n)$ and $\hat{\mathbb{E}}_n\|W_1\|^3 = \frac{1}{n}\sum_{i=1}^n W_i^3$.

By the strong law of large number, $\hat{\mathbb{E}}_n\|W_1\|^3 < 2\mathbb{E}\|W_1\|^3$ almost surely. Because $\mathsf{Var}(W_1^*|\mathcal{X}_n) - \mathsf{Var}(W_1) = O_P(\frac{1}{\sqrt{n}})$ implies $\sup_t |P(Z_w^* < t|\mathcal{X}_n) - P(Z_w < t)| = O_P(\frac{1}{\sqrt{n}})$, we conclude

$$\sup_t \left|P\left(a^T\Delta_n^* < t|\mathcal{X}_n\right) - P\left(a^T\Delta_n < t\right)\right| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

Finally, by choosing a to be the unit vector along each coordinate, we obtain the desire result for the bootstrap CIs. □

PROOF OF THEOREM 3. The high level ideas of this proof is very similar to that of the previous theorem. In the first part, we derive the Berry–Esseen bound of the ELBO estimator. In the second part, we prove the bootstrap consistency. Note that in the increasing-dimensional case, many smaller-order approximations (e.g., those from a Taylor expansion) may depend on the dimension d and may no longer be small. So we need to examine each approximation term.

Part I: Berry–Esseen Bound. Recall that $\Psi(\theta) = \mathbb{E}(\mathcal{E}(\theta|X_1))$ and Ψ_θ, $\Psi_{\theta\theta}$ are the gradient and Hessian matrix of Ψ. Let

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^n \mathcal{E}(\theta|X_i),$$

$$\Psi_{\theta,n}(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla\mathcal{E}(\theta|X_i),$$

$$\Psi_{\theta\theta,n}(\theta) = \frac{1}{n}\sum_{i=1}^n \nabla\nabla\mathcal{E}(\theta|X_i)$$

denote the corresponding empirical versions.

Because $0 = \Psi_{\theta,n}(\hat{\theta}_{\mathrm{ELBO}}) = \Psi_\theta(\theta_{\mathrm{ELBO}})$, by Taylor's theorem

$$\begin{aligned}\Psi_{\theta,n}(\theta_{\mathrm{ELBO}}) - \Psi_\theta(\theta_{\mathrm{ELBO}}) &= \Psi_{\theta,n}(\theta_{\mathrm{ELBO}}) - \Psi_{\theta,n}(\hat{\theta}_{\mathrm{ELBO}}) \\ &= \Psi_{\theta\theta,n}(\theta_{\mathrm{ELBO}})(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) + E_{1,n} \\ &= \left(\Psi_{\theta\theta}(\theta_{\mathrm{ELBO}}) + E_{2,n}\right)(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) + E_{1,n}\end{aligned}$$

$$= \Psi_{\theta\theta}(\theta_{\mathrm{ELBO}})(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) + E_{2,n}(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) + E_{1,n},$$

where $E_{1,n} \in \mathbb{R}^d$ is a vector about the second order Taylor approximation error and $E_{2,n} = \Psi_{\theta\theta,n}(\theta_{\mathrm{ELBO}}) - \Psi_{\theta\theta}(\theta_{\mathrm{ELBO}})$.

Let $Z_n = \Psi_{\theta,n}(\theta_{\mathrm{ELBO}}) - \Psi_{\theta}(\theta_{\mathrm{ELBO}})$ denotes the empirical gradient minus the corresponding expected gradient. By assumption (A1), $\Psi_{\theta\theta}(\theta_{\mathrm{ELBO}})$ is always invertible, so multiplying $\Omega = \Psi_{\theta\theta}^{-1}(\theta_{\mathrm{ELBO}})$ in both sides and rearranging the equation lead to

$$\begin{aligned}\tilde{\Delta}_n &= \hat{\theta}_{\mathrm{ELBO}} - \theta_{\mathrm{ELBO}} \\ &= -\Omega Z_n - \Omega E_{2,n}(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) - \Omega E_{1,n}.\end{aligned}$$

The first quantity ΩZ_n has an asymptotic normality because it contains an empirical sum minus the corresponding expectation.

To derive the asymptotic normality of Δ_n, we need

$$\begin{aligned}&\|\sqrt{n}a_n^T\tilde{\Delta}_n + \sqrt{n}a_n^T\Omega Z_n\| \\ &\quad= \sqrt{n}\|\underbrace{a_n^T\Omega E_{2,n}(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})}_{\text{(I)}} - \underbrace{a_n^T\Omega E_{1,n}}_{\text{(II)}}\| \\ &\quad= o_P(1)\end{aligned} \tag{18}$$

for any sequence of unit vectors $a_n \in \mathbb{R}^d$.

For part (I), because $E_{2,n} = \Psi_{\theta\theta,n}(\theta_{\mathrm{ELBO}}) - \Psi_{\theta\theta}(\theta_{\mathrm{ELBO}})$ is the average of IID random matrices minus the corresponding expectation, the matrix Bernstein inequality [see, e.g., Theorem 6.2 in Tropp (2012)] implies $\|E_{2,n}\|_2 = O_P(\sqrt{\frac{\log^2 d}{n}})$, where $\|\cdot\|_2$ is the matrix 2-norm. This, along with the fact that assumption (A1) implies $\|\Omega\|_2$ being bounded, implies

$$\begin{aligned}&\sqrt{n}\|a_n^T\Omega E_{2,n}(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})\| \\ &\quad\le \sqrt{n}\|a_n\|\|\Omega\|_2 \underbrace{\|E_{2,n}\|_2}_{=O_P(\sqrt{\frac{\log^2 d}{n}})} \underbrace{\|\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}\|}_{=O_P(\sqrt{\frac{d}{n}})} \\ &\quad= O_P\left(\sqrt{\frac{d\log^2 d}{n}}\right).\end{aligned}$$

This bounds the contribution of (I).

For part (II), we only need to focus on bounding $\|E_{1,n}\|$ because $\|\Omega\|_2$ is bounded. By the Taylor's theorem, the ℓth element of $E_{1,n}$ can be written as

$$E_{1,n,\ell} = (\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})^T A_\ell(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})$$

with

$$A_\ell = \int_{t=0}^{t=1} \frac{\partial}{\partial \theta_\ell} \Psi_{\theta\theta,n}\big(\hat{\theta}_{\mathrm{ELBO}} + t(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}})\big)\, dt.$$

Let $\hat{\mu} = \frac{\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}}{\|\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}\|}$ denote the direction of $\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}$, and $r_n = \|\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}\|$, and $e_\ell \in \mathbb{R}^d$ be the unit vector pointing toward the ℓth coordinate. Then we can rewrite $E_{1,n,\ell}$ as

$$E_{1,n,\ell} = r_n^2 \int_{t=0}^{t=1} \nabla_{e_\ell} \nabla_{\hat{\mu}} \nabla_{\hat{\mu}} \Psi_n(\hat{\theta}_{\mathrm{ELBO}} + t \cdot r_n \cdot \hat{\mu}_n)\, dt.$$

Therefore,

$$E_{1,n} = r_n^2 \int_{t=0}^{t=1} \nabla \nabla_{\hat{\mu}} \nabla_{\hat{\mu}} \Psi_n(\hat{\theta}_{\mathrm{ELBO}} + t \cdot r_n \cdot \hat{\mu}_n)\, dt$$

and assumption (A2) implies that

$$\begin{aligned} \|E_{1,n}\| &= r_n^2 \left\| \int_{t=0}^{t=1} \nabla \nabla_{\hat{\mu}} \nabla_{\hat{\mu}} \Psi_n(\hat{\theta}_{\mathrm{ELBO}} + t \cdot r_n \cdot \hat{\mu}_n)\, dt \right\| \\ &\le r_n^2 c_1 \\ &= O_P\left(\frac{d}{n}\right). \end{aligned} \tag{19}$$

By assumption (A1), $\|\Omega\|_2$ bounded so

$$\|\sqrt{n} a_n^T \Omega E_{1,n}\| \le \sqrt{n} \|\Omega\|_2 \|E_{1,n}\| = O_P\left(\sqrt{\frac{d^2}{n}}\right),$$

which bounds (II).

As a result, the assumption $\frac{d^2}{n} \to 0$ implies

$$\sqrt{n} \|a_n^T \Omega E_{2,n}(\theta_{\mathrm{ELBO}} - \hat{\theta}_{\mathrm{ELBO}}) - a_n^T \Omega E_{1,n}\| = o_P(1)$$

so equation (18) holds.

To obtain the Berry–Esseen bound, after rearranging equation (18),

$$\sqrt{n} a_n^T \tilde{\Delta}_n = -\sqrt{n} a_n^T \Omega Z_n + o_P(1) = \sqrt{n} \bar{W}_n + o_P(1),$$

where $\bar{W}_n = \frac{1}{n} \sum_{i=1}^n W_i$ and $W_i = -a_n^T \Omega(\Psi_\theta(\theta_{\mathrm{ELBO}}|X_i) - \Psi_\theta(\theta_{\mathrm{ELBO}}))$. Note that the $W_1, \ldots, W_n$ are also IID. Thus, by assumption (A3) and the Berry–Esseen theorem [Berry (1941), Esseen (1942)] we conclude that

$$\begin{aligned} \sup_t \big| P\big(\sqrt{n} a_n^T \tilde{\Delta}_n < t\big) - P\big(\sigma(a_n) Z < t\big) \big| &= o_P(1) + c_{\mathrm{BE}} \frac{\mathbb{E}(|a_n^T W_1|^3)}{\sqrt{n}} \\ &= o_P(1) + o(1), \end{aligned} \tag{20}$$

where $Z \sim N(0,1)$ and $\sigma^2(a_n) = \mathsf{Var}(W_1) = a_n^T \Omega \mathsf{Cov}(\Psi_\theta(\theta_{\mathrm{ELBO}}|X_1))\Omega a_n$.

Part II: Bootstrap. In the bootstrap world, we are sampling from $\hat{P}_n$. Thus, all the above derivations hold except that everything is conditional on $X_1, \ldots, X_n$ and the expectation is taken over $\hat{P}_n$ instead of P. So the derivation in part I leads to

$$
\begin{aligned}
&\sup_t \left| P\left(\sqrt{n} a_n^T \tilde{\Delta}_n^* < t \middle| \mathcal{X}_n\right) - P\left(\hat{\sigma}_n(a_n) Z < t \middle| \mathcal{X}_n\right)\right| \\
&\quad = o_P(1) + c_{\mathrm{BE}} \frac{\hat{\mathbb{E}}_n(|a_n^T W_1|^3)}{\sqrt{n}} \\
&\quad = o_P(1) + \frac{c_{\mathrm{BE}}}{\sqrt{n}} \cdot \frac{1}{n} \sum_{i=1}^n |a_n^T W_i|^3 \\
&\quad \le o_P(1) + \frac{c_{\mathrm{BE}}}{\sqrt{n}} \cdot \frac{1}{n} \sum_{i=1}^n \|W_i\|^3 \\
&\quad \le o_P(1) + \frac{c_{\mathrm{BE}}}{\sqrt{n}} \cdot \max\left\{\|W_1\|^3, \ldots, \|W_n\|^3\right\} \\
&\quad \le o_P(1) + O_P\left(\sqrt{\frac{\log d}{n}} \frac{d^{3/2}}{n^{3/2}}\right) = o_P(1),
\end{aligned}
\tag{21}
$$

where

$$
\begin{aligned}
\hat{\sigma}_n^2(a_n) &= a_n^T \hat{\Omega}_n \hat{\mathsf{Cov}}_n\left(\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_1)\right)\hat{\Omega}_n a_n, \\
\hat{\Omega}_n &= \Psi_{\theta\theta,n}^{-1}(\hat{\theta}_{\mathrm{ELBO}}), \\
\widehat{\mathsf{Cov}}_n\left(\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_1)\right) &= \sum_{i=1}^n \Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_i)\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_i)^T
\end{aligned}
$$

are the empirical versions of $\sigma^2(a_n)$, Ω and $\mathsf{Cov}(\Psi_\theta(\theta_{\mathrm{ELBO}}|X_1))$.

By matrix Bernstein inequality [Tropp (2012)], the difference

$$
\begin{aligned}
&\left|\hat{\sigma}_n(a_n) - \sigma(a_n)\right| \\
&\quad \le \sup_{a:\|a\|=1} \left|\hat{\sigma}_n(a) - \sigma(a)\right| \\
&\quad = \left\|\hat{\Omega}_n \widehat{\mathsf{Cov}}_n\left(\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_1)\right)\hat{\Omega}_n - \Omega \mathsf{Cov}\left(\Psi_\theta(\theta_{\mathrm{ELBO}}|X_1)\right)\Omega\right\|_2 \\
&\quad = O_P\left(\|\hat{\Omega}_n - \Omega\|_2 + \left\|\widehat{\mathsf{Cov}}_n\left(\Psi_\theta(\hat{\theta}_{\mathrm{ELBO}}|X_1)\right)\right.\right. \\
&\qquad \left.\left. - \mathsf{Cov}\left(\Psi_\theta(\theta_{\mathrm{ELBO}}|X_1)\right)\right\|_2\right) \\
&\quad = O_P\left(\sqrt{\frac{\log^2 d}{n}}\right) = o_P(1).
\end{aligned}
$$

Therefore,

$$\sup_t \left| P\left(\mathbb{P}\left(\sigma(a_n)Z < t\right) - P\left(\hat{\sigma}_n(a_n)Z < t | \mathcal{X}_n\right)\right| = o_P(1).$$

This, together with equations (20) and (21), implies

$$\sup_t \left| P\left(\sqrt{n}a_n^T \tilde{\Delta}_n^* < t | \mathcal{X}_n\right) - P\left(\sqrt{n}a_n^T \tilde{\Delta}_n < t\right)\right| = o_P(1).$$

Finally, by choosing a_n to be the unit vector along each coordinate, we obtain the desire result for the bootstrap CIs. □

APPENDIX B: ASSUMPTIONS IN WESTLING AND MCCORMICK (2015)

Here we describe the assumptions (B1)–(B5) in the Appendix of Westling and McCormick (2015).

(B1) For all $\theta \in \Theta$ and P_0-a.e. x, $\mathsf{ELBO}(\theta, \omega | x)$ is uniquely maximized at $\omega = \omega_{\max}(\theta | x)$, which is an element of Ω, an open subset of $\mathbb{R}^s$.

(B2) $\omega_{\max}(\theta | x)$ is a measurable function of x for all θ and twice continuously differentiable in a neighborhood of θ_{ELBO} for P_0-a.e. x.

(B3) $\mathsf{ELBO}(\theta, \omega | x)$ is twice continuously differentiable in a neighborhood of θ_{ELBO} and $\omega_{\max}(\theta_{\mathrm{ELBO}} | x)$ for P_0-a.e. x.

(B4) There exists $r_1 > 0$, $s(x) > 0$, $b_1(x)$ and $b_2(x)$ such that

1. For all $x \in \mathbb{R}^J$ and $\theta \in B(\theta_{\mathrm{ELBO}}, r_1)$,
$$\omega_{\max}(\theta | x) \in B\left(\omega_{\max}(\theta_{\mathrm{ELBO}} | x), s(x)\right).$$
2. For all $x \in \mathbb{R}^J$, $\theta_1, \theta_2 \in B(\theta_{\mathrm{ELBO}}, r_1)$ and $\omega_1, \omega_2 \in B(\omega_{\max}(\theta_{\mathrm{ELBO}} | x), s(x))$,
$$|\mathsf{ELBO}(\theta_1, \omega_1 | x) - \mathsf{ELBO}(\theta_2, \omega_2 | x)| \le b_1(x)\left(\|\theta_1 - \theta_2\| + \|\omega_1 - \omega_2\|\right).$$
3. For all $\theta_1, \theta_2 \in B(\theta_{\mathrm{ELBO}}, r_1)$,
$$\left\|\omega_{\max}(\theta_1 | x) - \omega_{\max}(\theta_2 | x)\right\| \le b_2(x)\|\theta_1 - \theta_2\|.$$
4. The functions $b_1, b_2 \in L_2(P_0)$.

(B5) $|\nabla^2 \mathcal{E}(\theta | x)| \le \kappa(x)$ for all θ in a neighborhood of θ_{ELBO} and P_0-a.e. x for an integrable function κ.

Acknowledgements. We thank the referee and the Editor for insightful comments. We also thank Fang Han for helpful comments on the theoretical results.

REFERENCES

AIROLDI, E., BLEI, D., XING, E. and FIENBERG, S. (2005). A latent mixed membership model for relational data. In *Proceedings of the* 3*rd International Workshop on Link Discovery* 82–89. ACM, New York.

AIROLDI, E. M., BLEI, D. M., FIENBERG, S. E. and XING, E. P. (2008). Mixed membership stochastic blockmodels. *J. Mach. Learn. Res.* **9** 1981–2014.

AIROLDI, E. M., BLEI, D. M., EROSHEVA, E. A. and FIENBERG, S. E. (2015). Introduction to mixed membership models and methods. In *Handbook of Mixed Membership Models and Their Applications*. *Chapman & Hall/CRC Handb. Mod. Stat. Methods* 3–13. CRC Press, Boca Raton, FL. MR3380022

ANDREWS, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

BABU, G. J. and SINGH, K. (1983). Inference on means using the bootstrap. *Ann. Statist.* **11** 999–1003. MR0707951

BERRY, A. C. (1941). The accuracy of the Gaussian approximation to the sum of independent variates. *Trans. Amer. Math. Soc.* **49** 122–136. MR0003498

BICKEL, P., CHOI, D., CHANG, X. and ZHANG, H. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *Ann. Statist.* **41** 1922–1943. MR3127853

BLEI, D. M. and JORDAN, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Anal.* **1** 121–143. MR2227367

BLEI, D. M., KUCUKELBIR, A. and MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *J. Amer. Statist. Assoc.* **112** 859–877. MR3671776

BLEI, D. M., NG, A. Y. and JORDAN, M. I. (2003). Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3** 993–1022.

BOX, G. E. P. (1976). Science and statistics. *J. Amer. Statist. Assoc.* **71** 791–799. MR0431440

CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. MR2988467

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

DAMIANOU, A., TITSIAS, M. K. and LAWRENCE, N. D. (2011). Variational Gaussian process dynamical systems. In *Advances in Neural Information Processing Systems* 2510–2518.

DAMIANOU, A. C., TITSIAS, M. K. and LAWRENCE, N. D. (2016). Variational inference for latent variables and uncertain inputs in Gaussian processes. *J. Mach. Learn. Res.* **17** Paper No. 42. MR3491136

DOUGLAS, J. (1997). Joint consistency of nonparametric item characteristic curve and ability estimation. *Psychometrika* **62** 7–28. MR1439472

EFRON, B. (1982a). *The Jackknife, the Bootstrap and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics* **38**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR0659849

EFRON, B. (1982b). Bootstrap methods: Another look at the jackknife. In *Breakthroughs in Statistics* 569–593.

EROSHEVA, E. A., FIENBERG, S. E. and JOUTARD, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *Ann. Appl. Stat.* **1** 502–537. MR2415745

EROSHEVA, E. A. and WHITE, T. (2006). Operational definition of chronic disability in the national long term care survey: Problems and suggestions. Working Paper.

ESSEEN, C.-G. (1942). On the Liapounoff limit of error in the theory of probability. *Ark. Mat. Astron. Fys.* **28A** 19. MR0011909

FAN, J. and ZHOU, W.-X. (2016). Guarding against spurious discoveries in high dimensions. *J. Mach. Learn. Res.* **17** Paper No. 203. MR3580356

GHAHRAMANI, Z. and BEAL, M. J. (2000). Variational inference for Bayesian mixtures of factor analysers. In *Advances in Neural Information Processing Systems* 449–455.

HABERMAN, S. J. (1977). Maximum likelihood estimates in exponential response models. *Ann. Statist.* **5** 815–841. MR0501540

HALL, P. (1992). *The Bootstrap and Edgeworth Expansion*. Springer, New York. MR1145237

HALL, P., ORMEROD, J. T. and WAND, M. P. (2011). Theory of Gaussian variational approximation for a Poisson mixed model. *Statist. Sinica* **21** 369–389. MR2796867

HALL, P., PHAM, T., WAND, M. P. and WANG, S. S. J. (2011). Asymptotic normality and valid inference for Gaussian variational approximation. *Ann. Statist.* **39** 2502–2532. MR2906876

HOROWITZ, J. L. (1997). Bootstrap methods in econometrics: Theory and numerical performance. *Econom. Soc. Monogr.* **28** 188–222.

JORDAN, M. I., GHAHRAMANI, Z., JAAKKOLA, T. S. and SAUL, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.* **37** 183–233.

KHAN, M. E., BOUCHARD, G., MURPHY, K. P. and MARLIN, B. M. (2010). Variational bounds for mixed-data factor analysis. In *Advances in Neural Information Processing Systems* 1108–1116.

KLAMI, A., VIRTANEN, S., LEPPÄAHO, E. and KASKI, S. (2015). Group factor analysis. *IEEE Trans. Neural Netw. Learn. Syst.* **26** 2136–2147. MR3453146

LATOUCHE, P., BIRMELÉ, E. and AMBROISE, C. (2012). Variational Bayesian inference and complexity control for stochastic block models. *Stat. Model.* **12** 93–115. MR2953099

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856

MAMMEN, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17** 382–400. MR0981457

MAMMEN, E. (1993). Bootstrap and wild bootstrap for high-dimensional linear models. *Ann. Statist.* **21** 255–285. MR1212176

MANTON, K. G., CORDER, L. S. and STALLARD, E. (1993). Estimates of change in chronic disability and institutional incidence and prevalence rates in the us elderly population from the 1982, 1984, and 1989 national long term care survey. *J. Gerontol.* **48** S153–S166.

NEYMAN, J. and SCOTT, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16** 1–32. MR0025113

O'HAGAN, A., MURPHY, T. B. and GORMLEY, I. C. (2015). On estimation of parameter uncertainty in model-based clustering. Preprint. Available at arXiv:1510.00551.

PORTNOY, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12** 1298–1309. MR0760690

PORTNOY, S. (1985). Asymptotic behavior of M estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13** 1403–1417. MR0811499

PORTNOY, S. (1988). Asymptotic behavior of likelihood methods for exponential families when the number of parameters tends to infinity. *Ann. Statist.* **16** 356–366. MR0924876

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

REDNER, R. A. and WALKER, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26** 195–239. MR0738930

SINGH, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Ann. Statist.* **9** 1187–1195. MR0630102

TROPP, J. A. (2012). User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.* **12** 389–434. MR2946459

VAN DER VAART, A. W. (1998). *Asymptotic Statistics*. *Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

WAINWRIGHT, M. J. and JORDAN, M. I. (2008). Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.* **1** 1–305.

WANG, Y. and BLEI, D. M. (2017). Frequentist consistency of variational Bayes. Preprint. Available at arXiv:1705.03439.

WANG, Y. S. and EROSHEVA, E. A. (2015). Fitting mixed membership models using mixedmem.

WANG, Y. S., MATSUEDA, R. L. and EROSHEVA, E. A. (2017). A variational EM method for mixed membership models with multivariate rank data: An analysis of public policy preferences. *Ann. Appl. Stat.* **11** 1452–1480. MR3709566

WASSERMAN, L. (2006). *All of Nonparametric Statistics*. Springer, New York. MR2172729

WASSERMAN, L., KOLAR, M. and RINALDO, A. (2013). Estimating undirected graphs under weak assumptions. Preprint. Available at arXiv:1309.6933.

WESTLING, T. and MCCORMICK, T. H. (2015). Establishing consistency and improving uncertainty estimates of variational inference through m-estimation. Preprint. Available at arXiv:1510.08151.

WOODBURY, M. A., CLIVE, J. and GARSON, A. (1978). Mathematical typology: A grade of membership technique for obtaining disease definition. *Comput. Biomed. Res.* **11** 277–298.

Y.-C. CHEN
Y. S. WANG
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195
USA
E-MAIL: yenchic@uw.edu
ysamwang@uw.edu

E. E. EROSHEVA
DEPARTMENT OF STATISTICS
SCHOOL OF SOCIAL WORK
AND
CENTER FOR STATISTICS AND THE SOCIAL SCIENCES
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON 98195
USA
E-MAIL: erosheva@uw.edu