Variational Inference: A Review for Statisticians

David M. Blei
Department of Computer Science and Statistics
Columbia University

Alp Kucukelbir
Department of Computer Science
Columbia University

Jon D. McAuliffe
Department of Statistics
University of California, Berkeley

August 9, 2016

Abstract

One of the core problems of modern statistics is to approximate difficult-to-compute probability distributions. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation involving the posterior distribution. In this paper, we review variational inference (VI), a method from machine learning that approximates probability distributions through optimization. VI has been used in many applications and tends to be faster than classical methods, such as Markov chain Monte Carlo sampling. The idea behind VI is to first posit a family of distributions and then to find the member of that family which is close to the target. Closeness is measured by Kullback-Leibler divergence. We review the ideas behind mean-field variational inference, discuss the special case of VI applied to exponential family models, present a full example with a Bayesian mixture of Gaussians, and derive a variant that uses stochastic optimization to scale up to massive data. We discuss modern research in VI and highlight important open problems. VI is powerful, but it is not yet well understood. Our hope in writing this paper is to catalyze statistical research on this class of algorithms.

Keywords: Algorithms; Statistical Computing; Computationally Intensive Methods.

1 Introduction

One of the core problems of modern statistics is to approximate difficult-to-compute probability distributions. This problem is especially important in Bayesian statistics, which frames all inference about unknown quantities as a calculation about the posterior. Modern Bayesian statistics relies on models for which the posterior is not easy to compute and corresponding algorithms for approximating them.

In this paper, we review variational inference (VI), a method from machine learning for approximating probability distributions (Jordan et al., 1999; Wainwright and Jordan, 2008). Variational inference is widely used to approximate posterior distributions for Bayesian models, an alternative strategy to Markov chain Monte Carlo (MCMC) sampling. Compared to MCMC, variational inference tends to be faster and easier to scale to large data—it has been applied to problems such as large-scale document analysis, computational neuroscience, and computer vision. But variational inference has been studied less rigorously than than MCMC, and its statistical properties are less well understood. In writing this paper, our hope is to catalyze statistical research on variational inference.

First, we set up the general problem. Consider a joint distribution of latent variables $\mathbf{z} = z_{1:m}$ and observations $\mathbf{x} = x_{1:n}$,

$$p(\mathbf{z}, \mathbf{x}) = p(\mathbf{z})p(\mathbf{x} \mid \mathbf{z}).$$

In Bayesian models, the latent variables help govern the distribution of the data. A Bayesian model draws the latent variables from a prior distribution $p(\mathbf{z})$ and then relates them to the observations through the likelihood $p(\mathbf{x} \mid \mathbf{z})$. Inference in a Bayesian model amounts to conditioning on data and computing the posterior $p(\mathbf{z} \mid \mathbf{x})$. In complex Bayesian models, this computation often requires approximate inference.

For decades, the dominant paradigm for approximate inference has been MCMC (Hastings, 1970; Gelfand and Smith, 1990). In MCMC, we first construct a Markov chain on $\mathbf{z}$ whose stationary distribution is the posterior $p(\mathbf{z} \mid \mathbf{x})$. Then, we sample from the chain for a long time to (hopefully) collect independent samples from the stationary distribution. Finally, we approximate the posterior with an empirical estimate constructed from the collected samples.

MCMC sampling has evolved into an indispensable tool to the modern Bayesian statistician. Landmark developments include the Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970), the Gibbs sampler (Geman and Geman, 1984) and its application to Bayesian statistics (Gelfand and Smith, 1990). MCMC algorithms are under active investigation. They have been widely studied, extended, and applied; see Robert and Casella (2004) for a perspective.

However, there are problems for which we cannot easily use this approach. These arise particularly when we need an approximate conditional faster than a simple MCMC algorithm can produce, such as when data sets are large or models are very complex. In these settings, variational inference provides a good alternative approach to approximate Bayesian inference.

Rather than use sampling, the main idea behind variational inference is to use optimization. First, we posit a *family* of approximate distributions $\mathcal{Q}$. This is a set of distributions over the latent variables. Then, we try to find the member of that family that minimizes the Kullback-Leibler (KL) divergence to the exact posterior,

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z}) \in \mathcal{Q}} \mathrm{KL}\left(q(\mathbf{z}) \| p(\mathbf{z} \mid \mathbf{x})\right). \tag{1}$$

Finally, we approximate the posterior with the optimized member of the family $q^*(\cdot)$.

Variational inference thus turns the inference problem into an optimization problem, and the reach of the family $\mathcal{Q}$ manages the complexity of this optimization. One of the key ideas behind variational inference is to choose $\mathcal{Q}$ to be flexible enough to capture a distribution close to $p(\mathbf{z} \mid \mathbf{x})$, but simple enough for efficient optimization.[1]

We emphasize that MCMC and variational inference are different approaches to solving the same problem. MCMC algorithms sample from a Markov chain; variational algorithms solve an optimization problem. MCMC algorithms approximate the posterior with samples from the chain; variational algorithms approximate the posterior with the result of the optimization.

Comparing variational inference and MCMC. When should a statistician use MCMC and when should she use variational inference? We will offer some guidance. MCMC methods tend to be more computationally intensive than variational inference but they also provide guarantees of producing (asymptotically) exact samples from the target distribution (Robert and Casella, 2004). Variational inference does not enjoy such guarantees—it can only find a distribution close to the target—but tends to be faster than MCMC. Because it rests on optimization, variational inference easily takes advantage of methods like stochastic optimization (Robbins and Monro, 1951; Kushner and Yin, 1997) and distributed optimization (though some MCMC methods can also exploit these innovations (Welling and Teh, 2011; Ahmed et al., 2012)).

Thus, variational inference is suited to large data sets and scenarios where we want to quickly explore many models; MCMC is suited to smaller data sets and scenarios where we happily pay a heavier computational cost for more precise samples. For example, we might use MCMC in a setting where we spent 20 years collecting a small but expensive data set, where we are confident that our model is appropriate, and where we require precise inferences. We might use variational inference when fitting a probabilistic model of text to one billion text documents and where the inferences will be used to serve search results to a large population of users. In this scenario, we can use distributed computation and stochastic optimization to scale and speed up inference, and we can easily explore many different models of the data.

The relative accuracy of variational inference and MCMC is still unknown. We do know that variational inference generally underestimates the variance of the posterior distribution; this is a consequence of its objective function. But, depending on the task at hand, underestimating the variance may be acceptable. Several lines of empirical research have shown that variational inference does not necessarily suffer in accuracy, e.g., in terms of posterior predictive distributions (Blei and Jordan, 2006; Braun and McAuliffe, 2010; Kucukelbir et al., 2016); other research focuses on where variational inference falls short, especially around the posterior variance, and tries to more closely match the inferences made by MCMC (Giordano et al., 2015). In general, a statistical theory and understanding around variational inference is an important open area of research (see Section 5.2). We can envision future results that outline which classes of models are particularly suited to each algorithm and perhaps even theory that bounds their accuracy. More broadly, variational inference is a valuable tool, alongside MCMC, in the statistician's toolbox.

Research on variational inference. The development of variational techniques for Bayesian inference followed two parallel, yet separate, tracks. Peterson and Anderson (1987) is arguably the first variational procedure for a particular model: a neural network. This paper, along with insights from statistical mechanics (Parisi, 1988), led to a flurry of variational inference procedures for a wide class of models (Saul et al., 1996; Jaakkola

[1] We focus here on KL$(q||p)$-based optimization, also called Kullback Leibler variational inference (Barber, 2012). Wainwright and Jordan (2008) emphasize that any procedure which uses optimization to approximate a distribution can be termed "variational inference." This includes methods that like expectation propagation (Minka, 2001), belief propagation (Yedidia et al., 2001), or even the Laplace approximation. We briefly discuss alternative divergence measures in Section 5.

and Jordan, 1996, 1997; Ghahramani and Jordan, 1997; Jordan et al., 1999). In parallel, Hinton and Van Camp (1993) proposed a variational algorithm for a similar neural network model. Neal and Hinton (1999) (first published in 1993) made important connections to the expectation maximization (EM) algorithm (Dempster et al., 1977), which then led to a variety of variational inference algorithms for other types of models (Waterhouse et al., 1996; MacKay, 1997).

Modern research on variational inference focuses on several aspects: tackling Bayesian inference problems that involve massive data; using improved optimization methods for solving Equation (1) (which is usually subject to local minima); developing generic variational inference, algorithms that are easy to apply to a wide class of models; and increasing the accuracy of variational inference, e.g., by stretching the boundaries of $\mathcal{Q}$ while managing complexity in optimization.

Organization of this paper. Section 2 describes the basic ideas behind the simplest approach to variational inference: mean-field inference and coordinate-ascent optimization. Section 3 works out the details for a Bayesian mixture of Gaussians, an example model familiar to many readers. Sections 4.1 and 4.2 describe variational inference for the class of models where the joint distribution of the latent and observed variables are in the exponential family—this includes many intractable models from modern Bayesian statistics and reveals deep connections between variational inference and the Gibbs sampler of Gelfand and Smith (1990). Section 4.3 expands on this algorithm to describe stochastic variational inference (Hoffman et al., 2013), which scales variational inference to massive data using stochastic optimization (Robbins and Monro, 1951). Finally, with these foundations in place, Section 5 gives a perspective on the field—applications in the research literature, a survey of theoretical results, and an overview of some open problems.

2 Variational inference

The goal of variational inference is to approximate a conditional distribution of latent variables given observed variables. The key idea is to solve this problem with optimization. We use a family of distributions over the latent variables, parameterized by free "variational parameters." The optimization finds the member of this family, i.e., the setting of the parameters, that is closest in KL divergence to the conditional of interest. The fitted variational distribution then serves as a proxy for the exact conditional distribution.

2.1 The problem of approximate inference

Let $\mathbf{x} = x_{1:n}$ be a set of observed variables and $\mathbf{z} = z_{1:m}$ be a set of latent variables, with joint distribution $p(\mathbf{z}, \mathbf{x})$. We omit constants, such as hyperparameters, from the notation.

The inference problem is to compute the conditional distribution of the latent variables given the observations, $p(\mathbf{z} \mid \mathbf{x})$. This conditional can be used to produce point or interval estimates of the latent variables, form predictive distributions of new data, and more.

We can write the conditional distribution as

$$p(\mathbf{z} \mid \mathbf{x}) = \frac{p(\mathbf{z}, \mathbf{x})}{p(\mathbf{x})}. \tag{2}$$

The denominator contains the marginal distribution of the observations, also called the *evidence*. We calculate it by marginalizing out the latent variables from the joint distribu-

tion,

$$p(\mathbf{x}) = \int p(\mathbf{z}, \mathbf{x}) \, \mathrm{d}\mathbf{z}. \tag{3}$$

For many models, this evidence integral is unavailable in closed form or requires exponential time to compute. The evidence is what we need to compute the conditional from the joint; this is why inference in such models is hard.

Note we assume that all unknown quantities of interest are represented as latent random variables. This includes parameters that might govern all the data, as found in Bayesian models, and latent variables that are "local" to individual data points. It might appear to the reader that variational inference is only relevant in Bayesian settings. It has certainly had a significant impact on applied Bayesian computation, and we will be focusing on Bayesian models here. We emphasize, however, that variational inference is a general-purpose tool for estimating conditional distributions. One need not be a Bayesian to have use for variational inference.

Bayesian mixture of Gaussians. Consider a Bayesian mixture of unit-variance univariate Gaussians. There are K mixture components, corresponding to K normal distributions with means $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_K\}$. The mean parameters are drawn independently from a common prior $p(\mu_k)$, which we assume to be a Gaussian $\mathcal{N}(0, \sigma^2)$; the prior variance σ^2 is a hyperparameter. To generate an observation x_i from the model, we first choose a cluster assignment c_i. It indicates which latent cluster x_i comes from and is drawn from a categorical distribution over $\{1, \ldots, K\}$. (We encode c_i as an indicator K-vector, all zeros except for a one in the position corresponding to x_i's cluster.) We then draw x_i from the corresponding Gaussian $\mathcal{N}(c_i^\top \boldsymbol{\mu}, 1)$.

The full hierarchical model is

$$\begin{aligned}
\mu_k &\sim \mathcal{N}(0, \sigma^2), & k &= 1, \ldots, K, & (4)\\
c_i &\sim \text{Categorical}(1/K, \ldots, 1/K), & i &= 1, \ldots, n, & (5)\\
x_i \mid c_i, \boldsymbol{\mu} &\sim \mathcal{N}\left(c_i^\top \boldsymbol{\mu}, 1\right) & i &= 1, \ldots, n. & (6)
\end{aligned}$$

For a sample of size n, the joint distribution of latent and observed variables is

$$p(\boldsymbol{\mu}, \mathbf{c}, \mathbf{x}) = p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(c_i) p(x_i \mid c_i, \boldsymbol{\mu}). \tag{7}$$

The latent variables are $\mathbf{z} = \{\boldsymbol{\mu}, \mathbf{c}\}$, the K class means and n class assignments.

Here, the evidence is

$$p(\mathbf{x}) = \int p(\boldsymbol{\mu}) \prod_{i=1}^{n} \sum_{c_i} p(c_i) p(x_i \mid c_i, \boldsymbol{\mu}) \, \mathrm{d}\boldsymbol{\mu}. \tag{8}$$

The integrand in Equation (8) does not contain a separate factor for each μ_k. (Indeed, each μ_k appears in all n factors of the integrand.) Thus, the integral in Equation (8) does not reduce to a product of one-dimensional integrals over the μ_k's. The time complexity of numerically evaluating the K-dimensional integral is $\mathcal{O}(K^n)$.

If we distribute the product over the sum in (8) and rearrange, we can write the evidence as a sum over all possible configurations $\mathbf{c}$ of cluster assignments,

$$p(\mathbf{x}) = \sum_{\mathbf{c}} p(\mathbf{c}) \int p(\boldsymbol{\mu}) \prod_{i=1}^{n} p(x_i \mid c_i, \boldsymbol{\mu}) \, \mathrm{d}\boldsymbol{\mu}. \tag{9}$$

Here each individual integral is computable, thanks to the conjugacy between the Gaussian prior on the components and the Gaussian likelihood. But there are K^n of them, one for each configuration of the cluster assignments. Computing the evidence remains exponential in K, hence intractable.

2.2 The evidence lower bound

In variational inference, we specify a family $\mathcal{Q}$ of distributions over the latent variables. Each $q(\mathbf{z}) \in \mathcal{Q}$ is a candidate approximation to the exact conditional. Our goal is to find the best candidate, the one closest in KL divergence to the exact conditional.[2] Inference now amounts to solving the following optimization problem,

$$q^*(\mathbf{z}) = \arg\min_{q(\mathbf{z})\in\mathcal{Q}} \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z}\,|\,\mathbf{x})\right). \tag{10}$$

Once found, $q^*(\cdot)$ is the best approximation of the conditional, within the family $\mathcal{Q}$. The complexity of the family determines the complexity of this optimization.

However, this objective is not computable because it requires computing the evidence $\log p(\mathbf{x})$ in Equation (3). (That the evidence is hard to compute is why we appeal to approximate inference in the first place.) To see why, recall that KL divergence is

$$\text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z}\,|\,\mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z}\,|\,\mathbf{x})\right], \tag{11}$$

where all expectations are taken with respect to $q(\mathbf{z})$. Expand the conditional,

$$\text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z}\,|\,\mathbf{x})\right) = \mathbb{E}\left[\log q(\mathbf{z})\right] - \mathbb{E}\left[\log p(\mathbf{z},\mathbf{x})\right] + \log p(\mathbf{x}). \tag{12}$$

This reveals its dependence on $\log p(\mathbf{x})$.

Because we cannot compute the KL, we optimize an alternative objective that is equivalent to the KL up to an added constant,

$$\text{ELBO}(q) = \mathbb{E}\left[\log p(\mathbf{z},\mathbf{x})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right]. \tag{13}$$

This function is called the evidence lower bound (ELBO). The ELBO is the negative KL divergence of Equation (12) plus $\log p(\mathbf{x})$, which is a constant with respect to $q(\mathbf{z})$. Maximizing the ELBO is equivalent to minimizing the KL divergence.

Examining the ELBO gives intuitions about the optimal variational distribution. We rewrite the ELBO as a sum of the expected log likelihood of the data and the KL divergence between the prior $p(\mathbf{z})$ and $q(\mathbf{z})$,

$$\begin{aligned}\text{ELBO}(q) &= \mathbb{E}\left[\log p(\mathbf{z})\right] + \mathbb{E}\left[\log p(\mathbf{x}\,|\,\mathbf{z})\right] - \mathbb{E}\left[\log q(\mathbf{z})\right] \\ &= \mathbb{E}\left[\log p(\mathbf{x}\,|\,\mathbf{z})\right] - \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z})\right).\end{aligned}$$

Which values of $\mathbf{z}$ will this objective encourage $q(\mathbf{z})$ to place its mass on? The first term is an expected likelihood; it encourages distributions that place their mass on configurations of the latent variables that explain the observed data. The second term is the negative divergence between the variational distribution and the prior; it encourages distributions close to the prior. Thus the variational objective mirrors the usual balance between likelihood and prior.

Another property of the ELBO is that it lower-bounds the (log) evidence, $\log p(\mathbf{x}) \geq \text{ELBO}(q)$ for any $q(\mathbf{z})$. This explains the name. To see this notice that Equations (12) and (13) give the following expression of the evidence,

$$\log p(\mathbf{x}) = \text{KL}\left(q(\mathbf{z}) \| p(\mathbf{z}\,|\,\mathbf{x})\right) + \text{ELBO}(q). \tag{14}$$

[2] The KL divergence is an information-theoretic measure of proximity between two distributions. It is asymmetric—that is, $\text{KL}\left(q\|p\right) \neq \text{KL}\left(p\|q\right)$—and nonnegative. It is minimized when $q(\cdot) = p(\cdot)$.

The bound then follows from the fact that $\text{KL}(\cdot) \geq 0$ (Kullback and Leibler, 1951). In the original literature on variational inference, this was derived through Jensen's inequality (Jordan et al., 1999).

The relationship between the ELBO and $\log p(\mathbf{x})$ has led to using the variational bound as a model selection criterion. This has been explored for mixture models (Ueda and Ghahramani, 2002; McGrory and Titterington, 2007) and more generally (Beal and Ghahramani, 2003). The premise is that the bound is a good approximation of the marginal likelihood, which provides a basis for selecting a model. Though this sometimes works in practice, selecting based on a bound is not justified in theory. Other research has used variational approximations in the log predictive density to use VI in cross-validation based model selection (Nott et al., 2012).

Finally, many readers will notice that the first term of the ELBO in Equation (13) is the expected complete log-likelihood, which is optimized by the EM algorithm (Dempster et al., 1977). The EM algorithm was designed for finding maximum likelihood estimates in models with latent variables. It uses the fact that the ELBO is equal to the log likelihood $\log p(\mathbf{x})$ (i.e., the log evidence) when $q(\mathbf{z}) = p(\mathbf{z} \mid \mathbf{x})$. EM alternates between computing the expected complete log likelihood according to $p(\mathbf{z} \mid \mathbf{x})$ (the E step) and optimizing it with respect to the model parameters (the M step). Unlike variational inference, EM assumes the expectation under $p(\mathbf{z} \mid \mathbf{x})$ is computable and uses it in otherwise difficult parameter estimation problems. Unlike EM, variational inference does not estimate fixed model parameters—it is often used in a Bayesian setting where classical parameters are treated as latent variables. Variational inference applies to models where we cannot compute the exact conditional of the latent variables.[3]

2.3 The mean-field variational family

We described the ELBO, the variational objective function in the optimization of Equation (10). We now describe a variational family $\mathcal{Q}$, to complete the specification of the optimization problem. The complexity of the family determines the complexity of the optimization; it is more difficult to optimize over a complex family than a simple family.

In this review we focus on the *mean-field variational family*, where the latent variables are mutually independent and each governed by a distinct factor in the variational distribution. A generic member of the mean-field variational family is

$$q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j). \tag{15}$$

Each latent variable z_j is governed by its own distribution $q_j(z_j)$. In optimization, these variational factors are chosen to maximize the ELBO of Equation (13).

We emphasize that the variational family is not a model of the observed data—indeed, the data $\mathbf{x}$ does not appear in Equation (15). Instead, it is the ELBO, and the corresponding KL minimization problem, that connects the fitted variational distribution to the data and model.

Notice we have not specified the parametric form of the individual variational factors. In principle, each can take on any parametric form appropriate to the corresponding random variable. For example, a continuous variable might have a Gaussian factor; a categorical variable will typically have a categorical factor. We will see in Sections 4 and 4.2 that

[3]Two notes: (a) Variational EM is the EM algorithm with a variational E-step, i.e., a computation of an approximate conditional. (b) The coordinate ascent algorithm of Section 2.4 can look like the EM algorithm. The "E step" computes approximate conditionals of local latent variables; the "M step" computes a conditional of the global latent variables.

there are many models for which properties of the model determine optimal forms of the mean-field factors.

Finally, though we focus on mean-field inference in this review, researchers have also studied more complex families. One way to expand the family is to add dependencies between the variables (Saul and Jordan, 1996; Barber and Wiegerinck, 1999); this is called structured variational inference. Another way to expand the family is to consider mixtures of variational distributions, i.e., additional latent variables within the variational family (Bishop et al., 1998). Both of these methods potentially improve the fidelity of the approximation, but there is a trade off. Structured and mixture-based variational families come with a more difficult-to-solve variational optimization problem.

Bayesian mixture of Gaussians (continued). Consider again the Bayesian mixture of Gaussians. The mean-field variational family contains approximate posterior distributions of the form

$$q(\boldsymbol{\mu}, \mathbf{c}) = \prod_{k=1}^{K} q(\mu_k; m_k, s_k^2) \prod_{i=1}^{n} q(c_i; \varphi_i). \tag{16}$$

Following the mean-field recipe, each latent variable is governed by its own variational factor. The factor $q(\mu_k; m_k, s_k^2)$ is a Gaussian distribution on the kth mixture component's mean parameter; its mean is m_k and its variance is s_k^2. The factor $q(c_i; \varphi_i)$ is a distribution on the ith observation's mixture assignment; its assignment probabilities are a K-vector φ_i.

Here we have asserted parametric forms for these factors: the mixture components are Gaussian with variational parameters (mean and variance) specific to the kth cluster; the cluster assignments are categorical with variational parameters (cluster probabilities) specific to the ith data point. In fact, these are the optimal forms of the mean-field variational distribution for the mixture of Gaussians.

With the variational family in place, we have completely specified the variational inference problem for the mixture of Gaussians. The ELBO is defined by the model definition in Equation (7) and the mean-field family in Equation (16). The corresponding variational optimization problem maximizes the ELBO with respect to the variational parameters, i.e., the Gaussian parameters for each mixture component and the categorical parameters for each cluster assignment. We will see this example through in Section 3.

Visualizing the mean-field approximation. The mean-field family is expressive because it can capture any marginal distribution of the latent variables. However, it cannot capture correlation between them. Seeing this in action reveals some of the intuitions and limitations of mean-field variational inference.

Consider a two dimensional Gaussian distribution, shown in violet in Figure 1. This distribution is highly correlated, which defines its elongated shape.

The optimal mean-field variational approximation to this posterior is a product of two Gaussian distributions. Figure 1 shows the mean-field variational density after maximizing the ELBO. While the variational approximation has the same mean as the original distribution, its covariance structure is, by construction, decoupled.

Further, the marginal variances of the approximation under-represent those of the target distribution. This is a common effect in mean-field variational inference and, with this example, we can see why. The KL divergence from the approximation to the posterior is in Equation (11). It penalizes placing mass in $q(\cdot)$ on areas where $p(\cdot)$ has little mass, but penalizes less the reverse. In this example, in order to successfully match the marginal variances, the circular $q(\cdot)$ would have to expand into territory where $p(\cdot)$ has little mass.

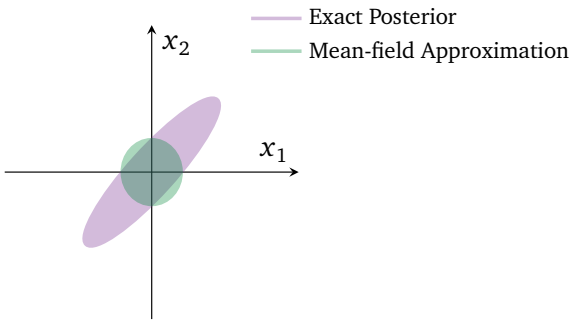

Figure 1: Visualizing the mean-field approximation to a two-dimensional Gaussian posterior. The ellipses (2σ) show the effect of mean-field factorization.

2.4 Coordinate ascent mean-field variational inference

Using the ELBO and the mean-field family, we have cast approximate conditional inference as an optimization problem. In this section, we describe one of the most commonly used algorithms for solving this optimization problem, coordinate ascent variational inference (CAVI) (Bishop, 2006). CAVI iteratively optimizes each factor of the mean-field variational distribution, while holding the others fixed. It climbs the ELBO to a local optimum.

The algorithm. We first state a result. Consider the jth latent variable z_j. The *complete conditional* of z_j is its conditional distribution given all of the other latent variables in the model and the observations, $p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})$. Fix the other variational factors $q_\ell(z_\ell)$, $\ell \neq j$. The optimal $q_j(z_j)$ is then proportional to the exponentiated expected log of the complete conditional,

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\}. \tag{17}$$

The expectation in Equation (17) is with respect to the (currently fixed) variational distribution over $\mathbf{z}_{-j}$, that is, $\prod_{\ell \neq j} q_\ell(z_\ell)$. Equivalently, Equation (17) is proportional to the exponentiated log of the joint,

$$q_j^*(z_j) \propto \exp\left\{\mathbb{E}_{-j}\left[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})\right]\right\}. \tag{18}$$

Because of the mean-field family—that all the latent variables are independent—the expectations on the right hand side do not involve the jth variational factor. Thus this is a valid coordinate update.

These equations underlie the CAVI algorithm, presented as Algorithm 1. We maintain a set of variational factors $q_\ell(z_\ell)$. We iterate through them, updating $q_j(z_j)$ using Equation (18). CAVI goes uphill on the ELBO of Equation (13), eventually finding a local optimum. As examples we show CAVI for a mixture of Gaussians in Section 3 and for a nonconjugate linear regression in Appendix A.

CAVI can also be seen as a "message passing" algorithm (Winn and Bishop, 2005), iteratively updating each random variable's variational parameters based on the variational parameters of the variables in its Markov blanket. This perspective enabled the design of automated software for a large class of models (Wand et al., 2011; Minka et al., 2014). Variational message passing connects variational inference to the classical theories of graphical models and probabilistic inference (Pearl, 1988; Lauritzen and Spiegelhalter, 1988). It has been extended to nonconjugate models (Knowles and Minka, 2011) and generalized via factor graphs (Minka, 2005).

Finally, CAVI is closely related to Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990), the classical workhorse of approximate inference. The Gibbs sampler

Algorithm 1: Coordinate ascent variational inference (CAVI)

Input: A model $p(\mathbf{x}, \mathbf{z})$, a data set $\mathbf{x}$

Output: A variational distribution $q(\mathbf{z}) = \prod_{j=1}^{m} q_j(z_j)$

Initialize: Variational factors $q_j(z_j)$

while *the* ELBO *has not converged* **do**

  for $j \in \{1, \ldots, m\}$ **do**

    Set $q_j(z_j) \propto \exp\{\mathbb{E}_{-j}[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})]\}$

  end

  Compute $\text{ELBO}(q) = \mathbb{E}\left[\log p(\mathbf{z}, \mathbf{x})\right] + \mathbb{E}\left[\log q(\mathbf{z})\right]$

end

return $q(\mathbf{z})$

maintains a realization of the latent variables and iteratively samples from each variable's complete conditional. Equation (18) uses the same complete conditional. It takes the expected log, and uses this quantity to iteratively set each variable's variational factor.[4]

Derivation. We now derive the coordinate update in Equation (18). The idea appears in Bishop (2006), but the argument there uses gradients, which we do not. Rewrite the ELBO of Equation (13) as a function of the jth variational factor $q_j(z_j)$, absorbing into a constant the terms that do not depend on it,

$$\text{ELBO}(q_j) = \mathbb{E}_j\left[\mathbb{E}_{-j}\left[\log p(z_j, \mathbf{z}_{-j}, \mathbf{x})\right]\right] - \mathbb{E}_j\left[\log q_j(z_j)\right] + \text{const.} \tag{19}$$

We have rewritten the first term of the ELBO using iterated expectation. The second term we have decomposed, using the independence of the variables (i.e., the mean-field assumption) and retaining only the term that depends on $q_j(z_j)$.

Up to an added constant, the objective function in Equation (19) is equal to the negative KL divergence between $q_j(z_j)$ and $q_j^*(z_j)$ from Equation (18). Thus we maximize the ELBO with respect to q_j when we set $q_j(z_j) = q_j^*(z_j)$.

2.5 Practicalities

Here, we highlight a few things to keep in mind when implementing and using variational inference in practice.

Initialization. The ELBO is (generally) a non-convex objective function. CAVI only guarantees convergence to a local optimum, which can be sensitive to initialization. Figure 2 shows the ELBO trajectory for 10 random initializations using the Gaussian mixture model. (This inference is on images; see Section 3.4.) Each initialization reaches a different value, indicating the presence of many local optima in the ELBO. In terms of $\text{KL}(q||p)$, better local optima give variational distributions that are closer to the exact posterior.

This is not always a disadvantage. Some models, such as the mixture of Gaussians (Section 3 and appendix B) and mixed-membership model (Appendix C), exhibit many posterior modes due to label switching: swapping cluster assignment labels induces many symmetric posterior modes. Representing one of these modes is sufficient for exploring latent clusters or predicting new observations.

[4] Many readers will know that we can significantly speed up the Gibbs sampler by marginalizing out some of the latent variables; this is called collapsed Gibbs sampling. We can speed up variational inference with similar reasoning; this is called collapsed variational inference. It has been developed for the same class of models described here (Sung et al., 2008; Hensman et al., 2012). These ideas are outside the scope of our review.

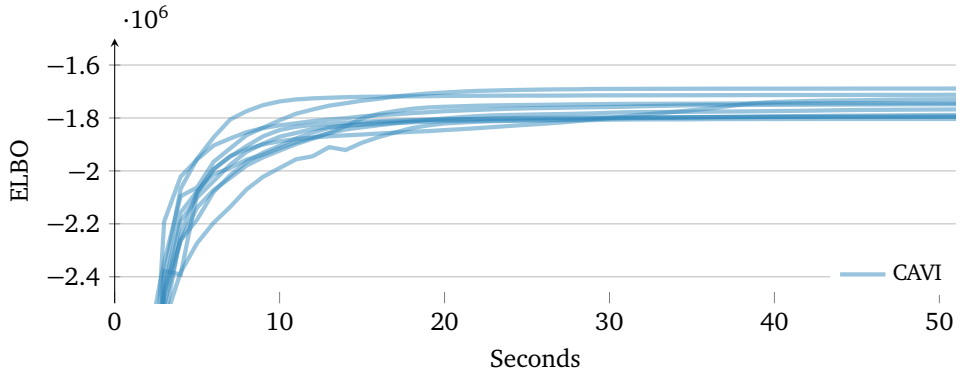

Figure 2: Different initializations may lead CAVI to find different local optima of the ELBO.

Assessing convergence. Monitoring the ELBO in CAVI is simple; we typically assess convergence once the change in ELBO has fallen below some small threshold. However, computing the ELBO of the full dataset may be undesirable. Instead, we suggest computing the average log predictive of a small held-out dataset. Monitoring changes here is a proxy to monitoring the ELBO of the full data. (Unlike the full ELBO, held-out predictive probability is not guaranteed to monotonically increase across iterations of CAVI.)

Numerical stability. Probabilities are constrained to live within $[0, 1]$. Precisely manipulating and performing arithmetic of small numbers requires additional care. When possible, we recommend working with logarithms of probabilities. One useful identity is the "log-sum-exp" trick,

$$\log\left[\sum_i \exp(x_i)\right] = a + \log\left[\sum_i \exp(x_i - a)\right]. \tag{20}$$

The constant a is typically set to $\max_i x_i$. This provides numerical stability to common computations in variational inference procedures.

3 A complete example: Bayesian mixture of Gaussians

As an example, we return to the simple mixture of Gaussians model of Section 2.1. To review, consider K mixture components and n real-valued data points $x_{1:n}$. The latent variables are K real-valued mean parameters $\boldsymbol{\mu} = \mu_{1:K}$ and n latent-class assignments $\mathbf{c} = c_{1:n}$. The assignment c_i indicates which latent cluster x_i comes from. In detail, c_i is an indicator K-vector, all zeros except for a one in the position corresponding to x_i's cluster. There is a fixed hyperparameter σ^2, the variance of the normal prior on the μ_k's. We assume the observation variance is one and take a uniform prior over the mixture components.

The joint distribution of the latent and observed variables is in Equation (7). The variational family is in Equation (16). Recall that there are two types of variational parameters—categorical parameters φ_i for approximating the posterior cluster assignment of the ith data point and Gaussian parameters m_k and s_k^2 for approximating the posterior of the kth mixture component.

We combine the joint and the mean-field family to form the ELBO for the mixture of

Gaussians. It is a function of the variational parameters $\mathbf{m}$, $\mathbf{s}^2$, and $\boldsymbol{\varphi}$,

$$\begin{aligned}\textsc{elbo}(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi}) = &\sum_{k=1}^{K} \mathbb{E}\left[\log p(\mu_k); m_k, s_k^2\right] \\ &+ \sum_{i=1}^{n} \left(\mathbb{E}\left[\log p(c_i); \varphi_i\right] + \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}, \mathbf{s}^2\right]\right) \\ &- \sum_{i=1}^{n} \mathbb{E}\left[\log q(c_i; \varphi_i)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q(\mu_k; m_k, s_k^2)\right].\end{aligned} \tag{21}$$

In each term, we have made explicit the dependence on the variational parameters. Each expectation can be computed in closed form.

The CAVI algorithm updates each variational parameter in turn. We first derive the update for the variational cluster assignment factor; we then derive the update for the variational mixture component factor.

3.1 The variational distribution of the mixture assignments

We first derive the variational update for the cluster assignment c_i. Using Equation (18),

$$q^*(c_i; \varphi_i) \propto \exp\left\{\log p(c_i) + \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \mathbf{m}, \mathbf{s}^2\right]\right\}. \tag{22}$$

The terms in the exponent are the components of the joint distribution that depend on c_i. The expectation in the second term is over the mixture components $\boldsymbol{\mu}$.

The first term of Equation (22) is the log prior of c_i. It is the same for all possible values of c_i, $\log p(c_i) = -\log K$. The second term is the expected log of the c_ith Gaussian density. Recalling that c_i is an indicator vector, we can write

$$p(x_i \mid c_i, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(x_i \mid \mu_k)^{c_{ik}}.$$

We use this to compute the expected log probability,

$$\begin{aligned}\mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu})\right] &= \sum_k c_{ik} \mathbb{E}\left[\log p(x_i \mid \mu_k); m_k, s_k^2\right] && (23)\\ &= \sum_k c_{ik} \mathbb{E}\left[-(x_i - \mu_k)^2; m_k, s_k^2\right] + \text{const.} && (24)\\ &= \sum_k c_{ik} \left(\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\right) + \text{const.} && (25)\end{aligned}$$

In each line we remove terms that are constant with respect to c_i. This calculation requires $\mathbb{E}\left[\mu_k\right]$ and $\mathbb{E}\left[\mu_k^2\right]$ for each mixture component, both computable from the variational Gaussian on the kth mixture component.

Thus the variational update for the ith cluster assignment is

$$\varphi_{ik} \propto \exp\left\{\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\right\}. \tag{26}$$

Notice it is only a function of the variational parameters for the mixture components.

3.2 The variational distribution of the mixture-component means

We turn to the variational distribution $q(\mu_k; m_k, s_k^2)$ of the kth mixture component. Again we use Equation (18) and write down the joint density up to a normalizing constant,

$$q(\mu_k) \propto \exp\left\{\log p(\mu_k) + \textstyle\sum_{i=1}^n \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right]\right\}. \tag{27}$$

We now calculate the unnormalized log of this coordinate-optimal $q(\mu_k)$. Recall φ_{ik} is the probability that the ith observation comes from the kth cluster. Because c_i is an indicator vector, we see that $\varphi_{ik} = \mathbb{E}\left[c_{ik}; \varphi_i\right]$. Now

$$\begin{aligned}
\log q(\mu_k) &= \log p(\mu_k) + \textstyle\sum_i \mathbb{E}\left[\log p(x_i \mid c_i, \boldsymbol{\mu}); \varphi_i, \mathbf{m}_{-k}, \mathbf{s}^2_{-k}\right] + \text{const.} && (28)\\
&= \log p(\mu_k) + \textstyle\sum_i \mathbb{E}\left[c_{ik} \log p(x_i \mid \mu_k); \varphi_i\right] + \text{const.} && (29)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \mathbb{E}\left[c_{ik}; \varphi_i\right] \log p(x_i \mid \mu_k) + \text{const.} && (30)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \varphi_{ik}\left(-(x_i - \mu_k)^2/2\right) + \text{const.} && (31)\\
&= -\mu_k^2/2\sigma^2 + \textstyle\sum_i \varphi_{ik} x_i \mu_k - \varphi_{ik}\mu_k^2/2 + \text{const.} && (32)\\
&= \left(\textstyle\sum_i \varphi_{ik} x_i\right)\mu_k - \left(1/2\sigma^2 + \textstyle\sum_i \varphi_{ik}/2\right)\mu_k^2 + \text{const.} && (33)
\end{aligned}$$

This calculation reveals that the coordinate-optimal variational distribution of μ_k is an exponential family with sufficient statistics $\{\mu_k, \mu_k^2\}$ and natural parameters $\{\sum_{i=1}^n \varphi_{ik} x_i, -1/2\sigma^2 - \sum_{i=1}^n \varphi_{ik}/2\}$, i.e., a Gaussian. Expressed in terms of the variational mean and variance, the updates for $q(\mu_k)$ are

$$m_k = \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}, \qquad s_k^2 = \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}. \tag{34}$$

These updates relate closely to the complete conditional distribution of the kth component in the mixture model. The complete conditional is a posterior Gaussian given the data assigned to the kth component. The variational update is a weighted complete conditional, where each data point is weighted by its variational probability of being assigned to component k.

3.3 CAVI for the mixture of Gaussians

Algorithm 2 presents coordinate-ascent variational inference for the Bayesian mixture of Gaussians. It combines the variational updates in Equation (22) and Equation (34). The algorithm requires computing the ELBO of Equation (21). We use the ELBO to track the progress of the algorithm and assess when it has converged.

Once we have a fitted variational distribution, we can use it as we would use the posterior. For example, we can obtain a posterior decomposition of the data. We assign points to their most likely mixture assignment $\hat{c}_i = \arg\max_k \varphi_{ik}$ and estimate cluster means with their variational means m_k.

We can also use the fitted variational distribution to approximate the predictive distribution of new data. This approximate predictive is a mixture of Gaussians,

$$p(x_{\text{new}} \mid x_{1:n}) \approx \frac{1}{K}\sum_{k=1}^K p(x_{\text{new}} \mid m_k), \tag{35}$$

where $p(x_{\text{new}} \mid m_k)$ is a Gaussian with mean m_k and unit variance.

Algorithm 2: CAVI for a Gaussian mixture model

Input: Data $x_{1:n}$, number of components K, prior variance of component means σ^2

Output: Variational distributions $q(\mu_k; m_k, s_k^2)$ (Gaussian) and $q(z_i; \varphi_i)$ (K-categorical)

Initialize: Variational parameters $\mathbf{m} = m_{1:K}$, $\mathbf{s}^2 = s_{1:K}^2$, and $\boldsymbol{\varphi} = \varphi_{1:n}$

while *the* ELBO *has not converged* **do**

  for $i \in \{1, \ldots, n\}$ **do**

    Set $\varphi_{ik} \propto \exp\{\mathbb{E}\left[\mu_k; m_k, s_k^2\right] x_i - \mathbb{E}\left[\mu_k^2; m_k, s_k^2\right]/2\}$

  end

  for $k \in \{1, \ldots, K\}$ **do**

$$\text{Set } m_k \longleftarrow \frac{\sum_i \varphi_{ik} x_i}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

$$\text{Set } s_k^2 \longleftarrow \frac{1}{1/\sigma^2 + \sum_i \varphi_{ik}}$$

  end

  Compute ELBO$(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

end

return $q(\mathbf{m}, \mathbf{s}^2, \boldsymbol{\varphi})$

3.4 Empirical study

We present two analyses to demonstrate the mixture of Gaussians algorithm in action. The first is a simulation study; the second is an analysis of a data set of natural images.

Simulation study. Consider two-dimensional real-valued data $\mathbf{x}$. We simulate $K = 5$ Gaussians with random means, covariances, and mixture assignments. Figure 3 shows the data; each point is colored according to its true cluster. Figure 3 also illustrates the initial variational distribution of the mixture components—each is a Gaussian, nearly centered, and with a wide variance; the subpanels plot the variational distribution of the components as the CAVI algorithm progresses.

The progression of the ELBO tells a story. We highlight key points where the ELBO develops "elbows", phases of the maximization where the variational approximation changes its shape. These "elbows" arise because the ELBO is not a convex function in terms of the variational parameters; CAVI iteratively reaches better plateaus.

Finally, we plot the logarithm of the Bayesian predictive distribution as approximated by the variational distribution. Here we report the average across held-out data. Note this plot is smoother than the ELBO.

Image analysis. We now turn to an experimental study. Consider the task of grouping images according to their color profiles. One approach is to compute the color histogram of the images. Figure 4 shows the red, green, and blue channel histograms of two images from the imageCLEF data (Villegas et al., 2013). Each histogram is a vector of length 192; concatenating the three color histograms gives a 576-dimensional representation of each image, regardless of its original size in pixel-space.

We use CAVI to fit a Gaussian mixture model to image histograms. We randomly select two sets of ten thousand images from the imageCLEF collection to serve as training and testing datasets. Figure 5 shows similarly colored images assigned to four randomly chosen clusters.

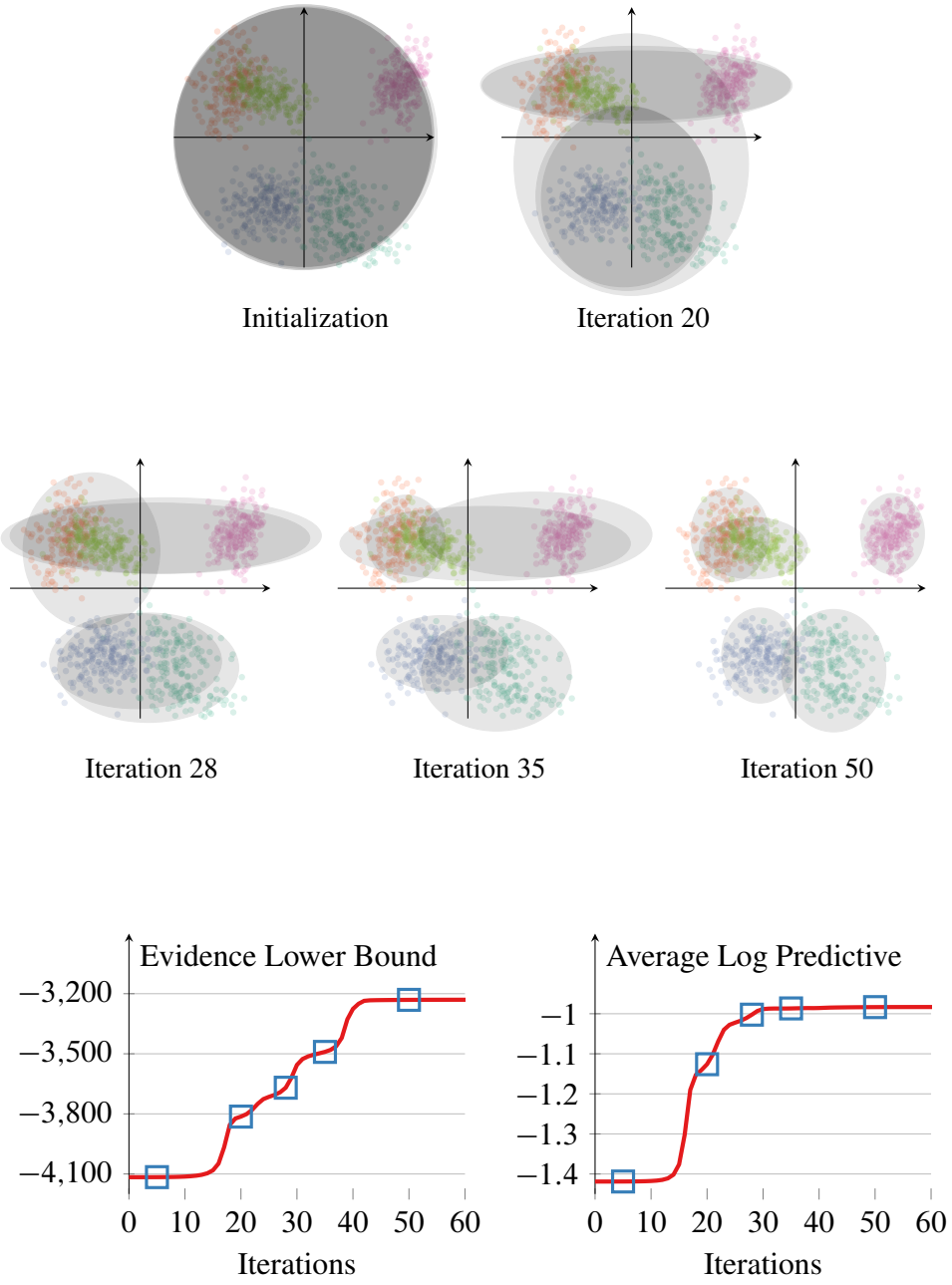

Figure 3: A simulation study of a two dimensional Gaussian mixture model. The ellipses are 2σ contours of the variational approximating factors.

Figure 6 shows the average log predictive accuracy of the testing set as a function of time. We compare CAVI to an implementation in Stan (Stan Development Team, 2015), which uses a Hamiltonian Monte Carlo-based sampler (Hoffman and Gelman, 2014). (Details are in Appendix B.) CAVI is orders of magnitude faster than this sampling algorithm.[5]

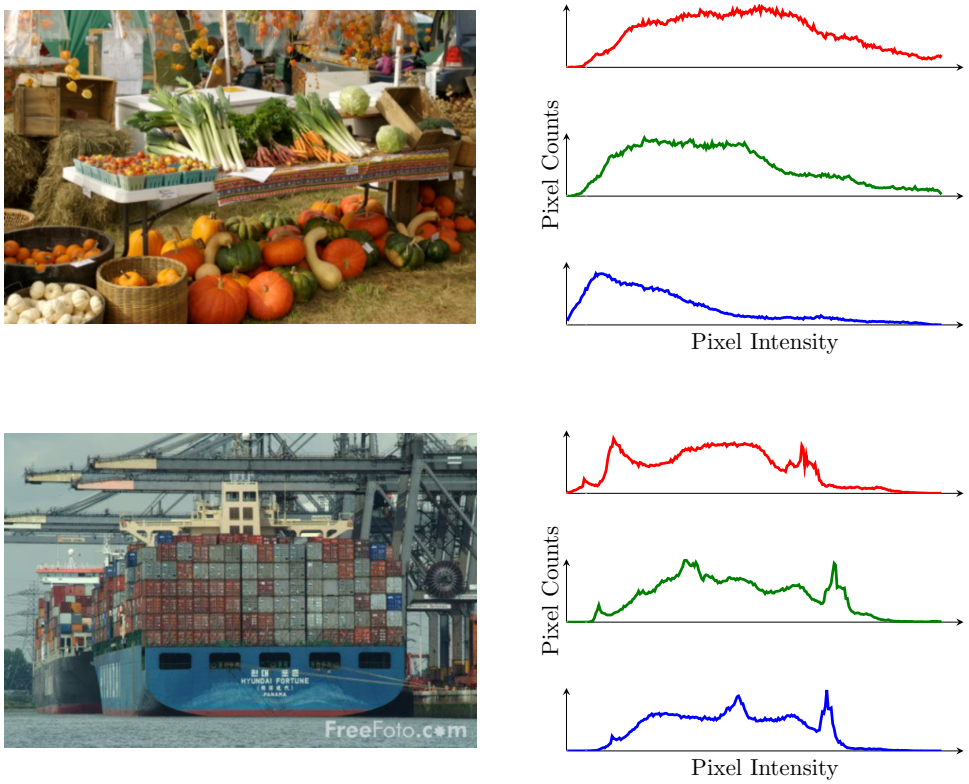

Figure 4: Red, green, and blue channel image histograms for two images from the imageCLEF dataset. The top image lacks blue hues, which is reflected in its blue channel histogram. The bottom image has a few dominant shades of blue and green, as seen in the peaks of its histogram.

4 Variational inference with exponential families

We described mean-field variational inference and derived CAVI, a general coordinate-ascent algorithm for optimizing the ELBO. We demonstrated this approach on a simple mixture of Gaussians, where each coordinate update was available in closed form.

The mixture of Gaussians is one member of the important class of models where each complete conditional is in the exponential family. This includes a number of widely used models, such as Bayesian mixtures of exponential families, factorial mixture models, matrix factorization models, certain hierarchical regression models (e.g., linear regression, probit regression, Poisson regression), stochastic blockmodels of networks, hierarchical mixtures of experts, and a variety of mixed-membership models (which we will discuss below).

Working in this family simplifies variational inference: it is easier to derive the corresponding CAVI algorithm, and it enables variational inference to scale up to massive data. In

[5]This is not a definitive comparison between variational inference and MCMC. Other samplers, such as a collapsed Gibbs sampler, may perform better than Hamiltonian Monte Carlo sampling.

(a) Purple **(b)** Green & White **(c)** Orange **(d)** Grayish Blue

Figure 5: Example clusters from the Gaussian mixture model. We assign each image to its most likely mixture cluster. The subfigures show nine randomly sampled images from four clusters; their namings are subjective.

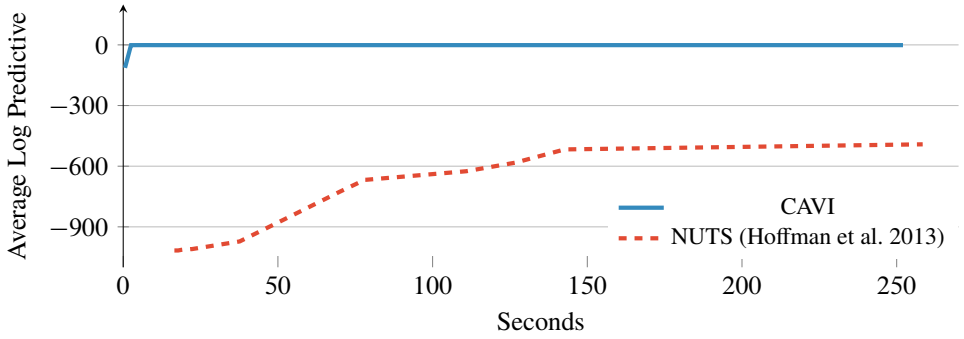

Figure 6: Comparison of CAVI to a Hamiltonian Monte Carlo-based sampling technique. CAVI fits a Gaussian mixture model to ten thousand images in less than a minute.

Section 4.1, we develop the general case. In Section 4.2, we discuss conditionally conjugate models, i.e., the common Bayesian application where some latent variables are "local" to a data point and others, usually identified with parameters, are "global" to the entire data set. Finally, in Section 4.3, we describe stochastic variational inference (Hoffman et al., 2013), a stochastic optimization algorithm that scales up variational inference in this setting.

4.1 Complete conditionals in the exponential family

Consider the generic model $p(\mathbf{z}, \mathbf{x})$ of Section 2.1 and suppose each complete conditional is in the exponential family:

$$p(z_j \mid \mathbf{z}_{-j}, \mathbf{x}) = h(z_j) \exp\{\eta_j(\mathbf{z}_{-j}, \mathbf{x})^\top z_j - a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\}, \tag{36}$$

where z_j is its own sufficient statistic, $h(\cdot)$ is a base measure, and $a(\cdot)$ is the log normalizer (Brown, 1986). Because this is a conditional distribution, the parameter $\eta_j(\mathbf{z}_{-j}, \mathbf{x})$ is a function of the conditioning set.

Consider mean-field variational inference for this class of models, where we fit $q(\mathbf{z}) = \prod_j q_j(z_j)$. The exponential family assumption simplifies the coordinate update of Equation (17),

$$q(z_j) \propto \exp\left\{\mathbb{E}\left[\log p(z_j \mid \mathbf{z}_{-j}, \mathbf{x})\right]\right\} \tag{37}$$

$$= \exp\left\{\log h(z_j) + \mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]^\top z_j - \mathbb{E}\left[a(\eta_j(\mathbf{z}_{-j}, \mathbf{x}))\right]\right\} \tag{38}$$

$$\propto h(z_j) \exp\left\{\mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]^\top z_j\right\}. \tag{39}$$

This update reveals the parametric form of the optimal variational factors. Each one is in the same exponential family as its corresponding complete conditional. Its parameter has the same dimension and it has the same base measure $h(\cdot)$ and log normalizer $a(\cdot)$.

Having established their parametric forms, let v_j denote the variational parameter for the jth variational factor. When we update each factor, we set its parameter equal to the expected parameter of the complete conditional,

$$v_j = \mathbb{E}\left[\eta_j(\mathbf{z}_{-j}, \mathbf{x})\right]. \tag{40}$$

This expression facilitates deriving CAVI algorithms for many complex models.

4.2 Conditional conjugacy and Bayesian models

One important special case of exponential family models are *conditionally conjugate models* with local and global variables. Models like this come up frequently in Bayesian statistics and statistical machine learning, where the global variables are the "parameters" and the local variables are per-data-point latent variables.

Conditionally conjugate models. Let β be a vector of *global latent variables*, which potentially govern any of the data. Let $\mathbf{z}$ be a vector of *local latent variables*, whose ith component only governs data in the ith "context." The joint distribution is

$$p(\beta, \mathbf{z}, \mathbf{x}) = p(\beta)\prod_{i=1}^{n} p(z_i, x_i \mid \beta). \tag{41}$$

The mixture of Gaussians of Section 3 is an example. The global variables are the mixture components; the ith local variable is the cluster assignment for data point x_i.

We will assume that the modeling terms of Equation (41) are chosen to ensure each complete conditional is in the exponential family. In detail, we first assume the joint distribution of each (x_i, z_i) pair, conditional on β, has an exponential family form,

$$p(z_i, x_i \mid \beta) = h(z_i, x_i)\exp\{\beta^\top t(z_i, x_i) - a(\beta)\}, \tag{42}$$

where $t(\cdot, \cdot)$ is the sufficient statistic.

Next, we take the prior on the global variables to be the corresponding conjugate prior (Diaconis et al., 1979; Bernardo and Smith, 1994),

$$p(\beta) = h(\beta)\exp\{\alpha^\top[\beta, -a(\beta)] - a(\alpha)\}. \tag{43}$$

This prior has natural (hyper)parameter $\alpha = [\alpha_1, \alpha_2]$ and sufficient statistics that concatenate the global variable and its log normalizer in the distribution of the local variables.

With the conjugate prior, the complete conditional of the global variables is in the same family. Its natural parameter is

$$\hat{\alpha} = \begin{bmatrix} \alpha_1 + \sum_{i=1}^{n} t(z_i, x_i) \\ \alpha_2 + n \end{bmatrix}. \tag{44}$$

Turn now to the complete conditional of the local variable z_i. Given β and x_i, the local variable z_i is conditionally independent of the other local variables $\mathbf{z}_{-i}$ and other data $\mathbf{x}_{-i}$. This follows from the form of the joint distribution in Equation (41). Thus

$$p(z_i \mid x_i, \beta, \mathbf{z}_{-i}, \mathbf{x}_{-i}) = p(z_i \mid x_i, \beta). \tag{45}$$

We further assume that this distribution is in an exponential family,

$$p(z_i \mid x_i, \beta) = h(z_i) \exp\{\eta(\beta, x_i)^\top z_i - a(\eta(\beta, x_i))\}. \tag{46}$$

This is a property of the local likelihood term $p(z_i, x_i \mid \beta)$ from Equation (42). For example, in the mixture of Gaussians, the complete conditional of the local variable is a categorical.

Variational inference in conditionally conjugate models. We now describe CAVI for this general class of models. Write $q(\beta \mid \lambda)$ for the variational posterior approximation on β; we call λ the "global variational parameter". It indexes the same exponential family distribution as the prior. Similarly, let the variational posterior $q(z_i \mid \varphi_i)$ on each local variable z_i be governed by a "local variational parameter" φ_i. It indexes the same exponential family distribution as the local complete conditional. CAVI iterates between updating each local variational parameter and updating the global variational parameter.

The local variational update is

$$\varphi_i = \mathbb{E}_\lambda \left[\eta(\beta, x_i)\right]. \tag{47}$$

This is an application of Equation (40), where we take the expectation of the natural parameter of the complete conditional in Equation (45).

The global variational update applies the same technique. It is

$$\lambda = \begin{bmatrix} \alpha_1 + \sum_{i=1}^n \mathbb{E}_{\varphi_i}\left[t(z_i, x_i)\right] \\ \alpha_2 + n \end{bmatrix}. \tag{48}$$

Here we take the expectation of the natural parameter in Equation (44).

CAVI optimizes the ELBO by iterating between local updates of each local parameter and global updates of the global parameters. To assess convergence we can compute the ELBO at each iteration (or at some lag), up to a constant that does not depend on the variational parameters,

$$\text{ELBO} = \left(\alpha_1 + \sum_{i=1}^n \mathbb{E}_{\varphi_i}\left[t(z_i, x_i)\right]\right)^\top \mathbb{E}_\lambda\left[\beta\right] - (\alpha_2 + n)\mathbb{E}_\lambda\left[a(\beta)\right] - \mathbb{E}\left[\log q(\beta, \mathbf{z})\right]. \tag{49}$$

This is the ELBO in Equation (13) applied to the joint in Equation (41) and the corresponding mean-field variational distribution; we have omitted terms that do not depend on the variational parameters. The last term is

$$\mathbb{E}\left[\log q(\beta, \mathbf{z})\right] = \lambda^\top \mathbb{E}_\lambda\left[t(\beta)\right] - a(\lambda) + \sum_{i=1}^n \varphi_i^\top \mathbb{E}_{\varphi_i}\left[z_i\right] - a(\varphi_i). \tag{50}$$

CAVI for the mixture of Gaussians model (Algorithm 2) is an instance of this method. Appendix C presents another example of CAVI for latent Dirichlet allocation (LDA), a probabilistic topic model.

4.3 Stochastic variational inference

Modern applications of probability models often require analyzing massive data. However, most posterior inference algorithms do not easily scale. CAVI is no exception, particularly in the conditionally conjugate setting of Section 4.2. The reason is that the coordinate ascent structure of the algorithm requires iterating through the entire data set at each iteration. As the data set size grows, each iteration becomes more computationally expensive.

An alternative to coordinate ascent is gradient-based optimization, which climbs the ELBO by computing and following its gradient at each iteration. This perspective is the key to scaling up variational inference using stochastic variational inference (SVI) (Hoffman et al., 2013), a method that combines natural gradients (Amari, 1998) and stochastic optimization (Robbins and Monro, 1951).

SVI focuses on optimizing the global variational parameters λ of a conditionally conjugate model. The flow of computation is simple. The algorithm maintains a current estimate of the global variational parameters. It repeatedly (a) subsamples a data point from the full data set; (b) uses the current global parameters to compute the optimal local parameters for the subsampled data point; and (c) adjusts the current global parameters in an appropriate way. SVI is detailed in Algorithm 3. We now show why it is a valid algorithm for optimizing the ELBO.

The natural gradient of the ELBO. In gradient-based optimization, the *natural gradient* accounts for the geometric structure of probability parameters (Amari, 1982, 1998). Specifically, natural gradients warp the parameter space in a sensible way, so that moving the same distance in different directions amounts to equal change in symmetrized KL divergence. The usual Euclidean gradient does not enjoy this property.

In exponential families, we find the natural gradient with respect to the parameter by premultiplying the usual gradient by the inverse covariance of the sufficient statistic, $a''(\lambda)^{-1}$. This is the inverse Riemannian metric and the inverse Fisher information matrix (Amari, 1982).

Conditionally conjugate models enjoy simple natural gradients of the ELBO. We focus on gradients with respect to the global parameter λ. Hoffman et al. (2013) derive the Euclidean gradient of the ELBO,

$$\nabla_\lambda \text{ELBO} = a''(\lambda)(\mathbb{E}_\varphi\left[\hat{\alpha}\right] - \lambda), \tag{51}$$

where $\mathbb{E}_\varphi\left[\hat{\alpha}\right]$ is in Equation (48). Premultiplying by the inverse Fisher information gives the natural gradient $g(\lambda)$,

$$g(\lambda) = \mathbb{E}_\varphi\left[\hat{\alpha}\right] - \lambda. \tag{52}$$

It is the difference between the coordinate updates $\mathbb{E}_\varphi\left[\hat{\alpha}\right]$ and the variational parameters λ at which we are evaluating the gradient. In addition to enjoying good theoretical properties, the natural gradient is easier to calculate than the Euclidean gradient. For more on natural gradients and variational inference see Sato (2001) and Honkela et al. (2008).

We can use this natural gradient in a gradient-based optimization algorithm. At each iteration, we update the global parameters,

$$\lambda_t = \lambda_{t-1} + \epsilon_t g(\lambda_t), \tag{53}$$

where ϵ_t is a step size.

Substituting Equation (52) into the second term reveals a special structure,

$$\lambda_t = (1 - \epsilon_t)\lambda_{t-1} + \epsilon_t \mathbb{E}_\varphi\left[\hat{\alpha}\right]. \tag{54}$$

Notice this does not require additional types of calculations other than those for coordinate ascent updates. At each iteration, we first compute the coordinate update. We then adjust the current estimate to be a weighted combination of the update and the current variational parameter.

Though easy to compute, using the natural gradient has the same cost as the coordinate update in Equation (48); it requires summing over the entire data set and computing

the optimal local variational parameters for each data point. With massive data, this is prohibitively expensive.

Stochastic optimization of the ELBO. Stochastic variational inference solves this problem by using the natural gradient in a stochastic optimization algorithm. Stochastic optimization algorithms follow noisy but cheap-to-compute gradients to reach the optimum of an objective function. (In the case of the ELBO, stochastic optimization will reach a local optimum.) In their seminal paper, Robbins and Monro (1951) proved results implying that optimization algorithms can successfully use noisy, unbiased gradients, as long as the step size sequence satisfies certain conditions. This idea has blossomed (Spall, 2003; Kushner and Yin, 1997). Stochastic optimization has enabled modern machine learning to scale to massive data (Le Cun and Bottou, 2004).

Our aim is to construct a cheaply computed, noisy, unbiased natural gradient. We expand the natural gradient in Equation (52) using Equation (44)

$$g(\lambda) = \alpha + \left[\sum_{i=1}^{n} \mathbb{E}_{\varphi_i^*}\left[t(z_i, x_i)\right], n\right] - \lambda, \tag{55}$$

where φ_i^* indicates that we consider the optimized local variational parameters (at fixed global parameters λ) in Equation (47). We construct a noisy natural gradient by sampling an index from the data and then rescaling the second term,

$$t \sim \text{Unif}(1, \ldots, n) \tag{56}$$

$$\hat{g}(\lambda) = \alpha + n\left[\mathbb{E}_{\varphi_t^*}\left[t(z_t, x_t)\right], 1\right] - \lambda. \tag{57}$$

The noisy natural gradient $\hat{g}(\lambda)$ is unbiased: $\mathbb{E}_t\left[\hat{g}(\lambda)\right] = g(\lambda)$. And it is cheap to compute—it only involves a single sampled data point and only one set of optimized local parameters. (This immediately extends to minibatches, where we sample B data points and rescale appropriately.) Again, the noisy gradient only requires calculations from the coordinate ascent algorithm. The first two terms of Equation (57) are equivalent to the coordinate update in a model with n replicates of the sampled data point.

Finally, we set the step size sequence. It must follow the conditions of Robbins and Monro (1951),

$$\sum_t \epsilon_t = \infty \quad ; \quad \sum_t \epsilon_t^2 < \infty. \tag{58}$$

Many sequences will satisfy these conditions, for example $\epsilon_t = t^{-\kappa}$ for $\kappa \in (0.5, 1]$. The full SVI algorithm is in Algorithm 3.

We emphasize that SVI requires no new derivation beyond what is needed for CAVI. Any implementation of CAVI can be immediately scaled up to a stochastic algorithm.

Probabilistic topic models. We demonstrate SVI with a probabilistic topic model. Probabilistic topic models are mixed-membership models of text, used to uncover the latent "topics" that run through a collection of documents. Topic models have become a popular technique for exploratory data analysis of large collections (Blei, 2012).

In detail, each latent topic is a distribution over terms in a vocabulary and each document is a collection of words that comes from a mixture of the topics. The topics are shared across the collection, but each document mixes them with different proportions. (This is the hallmark of a mixed-membership model.) Thus topic modeling casts topic discovery as a posterior inference problem. Posterior estimates of the topics and topic proportions can be used to summarize, visualize, explore, and form predictions about the documents.

One motivation for topic modeling is to get a handle on massive collections of documents. Early inference algorithms were based on coordinate ascent variational inference (Blei

Figure 7: Topics found in a corpus of 1.8M articles from the New York Times. Reproduced with permission from Hoffman et al. (2013).

et al., 2003) and analyzed collections in the thousands or tens of thousands of documents. (Appendix C presents this algorithm). With SVI, topic models scale up to millions of documents; the details of the algorithm are in Hoffman et al. (2013). Figure 7 illustrates topics inferred from 1.8M articles from the *New York Times*. This analysis would not have been possible without SVI.

5 Discussion

We described variational inference, a method that uses optimization to make probabilistic computations. The goal is to approximate the conditional distribution of latent variables $\mathbf{z}$ given observed variables $\mathbf{x}$, $p(\mathbf{z} \mid \mathbf{x})$. The idea is to posit a family of distributions $\mathcal{Q}$ and then to find the member $q^*(\cdot)$ that is closest in KL divergence to the conditional of interest. Minimizing the KL divergence is the optimization problem, and its complexity is governed by the complexity of the approximating family.

We then described the mean-field family, i.e., the family of fully factorized distributions of the latent variables. Using this family, variational inference is particularly amenable to coordinate-ascent optimization, which iteratively optimizes each factor. (This approach closely connects to the classical Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990).) We showed how to use mean-field VI to approximate the posterior distribution of a Bayesian mixture of Gaussians, discussed the special case of exponential families and conditional conjugacy, and described the extension to stochastic variational inference (Hoffman et al., 2013), which scales mean-field variational inference to massive data.

Algorithm 3: SVI for conditionally conjugate models

Input: Model $p(\mathbf{x}, \mathbf{z})$, data $\mathbf{x}$, and step size sequence ϵ_t

Output: Global variational distributions $q_\lambda(\beta)$

Initialize: Variational parameters λ_0

while *TRUE* **do**

Choose a data point uniformly at random, $t \sim \text{Unif}(1, \ldots, n)$

Optimize its local variational parameters $\varphi_t^* = \mathbb{E}_\lambda\left[\eta(\beta, x_t)\right]$

Compute the coordinate update as though x_t was repeated n times,

$$\hat{\lambda} = \alpha + n\mathbb{E}_{\varphi_t^*}\left[f(z_t, x_t)\right]$$

Update the global variational parameter, $\lambda_t = (1 - \epsilon_t)\lambda_t + \epsilon_t\hat{\lambda}_t$

end

return λ

5.1 Applications

Researchers in many fields have used variational inference to solve real problems. Here we focus on example applications of mean-field variational inference and structured variational inference based on the KL divergence. This discussion is not exhaustive; our intention is to outline the diversity of applications of variational inference.

Computational biology. VI is widely used in computational biology, where probabilistic models provide important building blocks for analyzing genetic data. For example, VI has been used in genome-wide association studies (Carbonetto and Stephens, 2012; Logsdon et al., 2010), regulatory network analysis (Sanguinetti et al., 2006), motif detection (Xing et al., 2004), phylogenetic hidden Markov models (Jojic et al., 2004), population genetics (Raj et al., 2014), and gene expression analysis (Stegle et al., 2010).

Computer vision and robotics. Since its inception, variational inference has been important to computer vision. Vision researchers frequently analyze large and high-dimensional data sets of images, and fast inference is important to successfully deploy a vision system. Some of the earliest examples included inferring non-linear image manifolds (Bishop and Winn, 2000) and finding layers of images in videos (Jojic and Frey, 2001). As other examples, variational inference is important to probabilistic models of videos (Chan and Vasconcelos, 2009; Wang and Mori, 2009), image denoising (Likas and Galatsanos, 2004), tracking (Vermaak et al., 2003; Yu and Wu, 2005), place recognition and mapping for robotics (Cummins and Newman, 2008; Ramos et al., 2012), and image segmentation with Bayesian nonparametrics (Sudderth and Jordan, 2009). Du et al. (2009) uses variational inference in a probabilistic model to combine the tasks of segmentation, clustering, and annotation.

Computational neuroscience. Modern neuroscience research also requires analyzing very large and high-dimensional data sets, such as high-frequency time series data or high-resolution functional magnetic imaging data. There have been many applications of variational inference to neuroscience, especially for autoregressive processes (Roberts and Penny, 2002; Penny et al., 2003, 2005; Flandin and Penny, 2007; Harrison and Green, 2010). Other applications of variational inference to neuroscience include hierarchical models of multiple subjects (Woolrich et al., 2004), spatial models (Sato et al., 2004; Zumer et al., 2007; Kiebel et al., 2008; Wipf and Nagarajan, 2009; Lashkari et al., 2012; Nathoo et al., 2014), brain-computer interfaces (Sykacek et al., 2004), and factor models (Manning et al., 2014; Gershman et al., 2014). There is a software toolbox that uses variational

methods for solving neuroscience and psychology research problems (Daunizeau et al., 2014).

Natural language processing and speech recognition. In natural language processing, variational inference has been used for solving problems such as parsing (Liang et al., 2007, 2009), grammar induction (Kurihara and Sato, 2006; Naseem et al., 2010; Cohen and Smith, 2010), models of streaming text (Yogatama et al., 2014), topic modeling (Blei et al., 2003), and hidden Markov models and part-of-speech tagging (Wang and Blunsom, 2013). In speech recognition, variational inference has been used to fit complex coupled hidden Markov models (Reyes-Gomez et al., 2004) and switching dynamic systems (Deng, 2004).

Other applications. There have been many other applications of variational inference. Fields in which it has been used include marketing (Braun and McAuliffe, 2010), optimal control and reinforcement learning (Van Den Broek et al., 2008; Furmston and Barber, 2010), statistical network analysis (Wiggins and Hofman, 2008; Airoldi et al., 2008), astrophysics (Regier et al., 2015), and the social sciences (Erosheva et al., 2007; Grimmer, 2011). General variational inference algorithms have been developed for a variety of classes of models, including shrinkage models (Armagan et al., 2011; Armagan and Dunson, 2011; Neville et al., 2014), general time-series models (Roberts et al., 2004; Barber and Chiappa, 2006; Archambeau et al., 2007b,a; Johnson and Willsky, 2014; Foti et al., 2014), robust models (Tipping and Lawrence, 2005; Wang and Blei, 2015), and Gaussian process models (Titsias and Lawrence, 2010; Damianou et al., 2011; Hensman et al., 2014).

5.2 Theory

Though researchers have not developed much theory around variational inference, there are several threads of research about theoretical guarantees of variational approximations. As we mentioned in the introduction, one of our purposes for writing this paper is to catalyze research on the statistical theory around variational inference.

Below, we summarize a variety of results. In general, they are all of the following type: treat VI posterior means as point estimates (or use M-step estimates from variational EM) and confirm that they have the usual frequentist asymptotics. (Sometimes the research finds that they do not enjoy the same asymptotics.) Each result revolves around a single model and a single family of variational approximations.

You et al. (2014) study the variational posterior for a classical Bayesian linear model. They put a normal prior on the coefficients and an inverse gamma prior on the response variance. They find that, under standard regularity conditions, the mean-field variational posterior mean of the parameters are consistent in the frequentist sense. Ormerod et al. (2014) build on their earlier work with a spike-and-slab prior on the coefficients and find similar consistency results.

Hall et al. (2011a,b) examine a simple Poisson mixed-effects model, one with a single predictor and a random intercept. They use a Gaussian variational approximation and estimate parameters with variational EM. They prove consistency of these estimates at the parametric rate and show asymptotic normality with asymptotically valid standard errors.

Celisse et al. (2012) and Bickel et al. (2013) analyze network data using stochastic blockmodels. They show asymptotic normality of parameter estimates obtained using a mean-field variational approximation. They highlight the computational advantages and theoretical guarantees of the variational approach over maximum likelihood for dense, sparse, and restricted variants of the stochastic blockmodel.

Westling and McCormick (2015) study the consistency of VI through a connection to M-estimation. They focus on a broader class of models (with posterior support in real coordinate space) and analyze an automated VI technique that uses a Gaussian variational approximation (Kucukelbir et al., 2015). They derive an asymptotic covariance matrix estimator and show its robustness to model misspecification.

Finally, Wang and Titterington (2006) analyze variational approximations to mixtures of Gaussians. Specifically, they consider Bayesian mixtures with conjugate priors, the mean-field variational approximation, and an estimator that is the variational posterior mean. They confirm that CAVI converges to a local optimum, that the VI estimator is consistent, and that the VI estimate and maximum likelihood estimate (MLE) approach each other at a rate of $\mathcal{O}(1/n)$. In Wang and Titterington (2005), they show that the asymptotic variational posterior covariance matrix is "too small"—it differs from the MLE covariance (i.e., the inverse Fisher information) by a positive-definite matrix.

5.3 Beyond conditional conjugacy

We focused on models where the complete conditional is in the exponential family. Many models, however, do not enjoy this property. A simple example is Bayesian logistic regression,

$$\beta_k \sim \mathcal{N}(0, 1),$$
$$y_i \mid x_i, \beta \sim \text{Bern}(\sigma(\beta^\top x_i)),$$

where $\sigma(\cdot)$ is the logistic function. The posterior distribution of the coefficients is not in an exponential family and we cannot apply the variational inference methods we discussed above. Specifically, we cannot compute the expectations in the first term of the ELBO in Equation (13) or the coordinate update in Equation (18).

Exploring variational methods for such models has been a fruitful area of research. An early example is Jaakkola and Jordan (1997, 2000), who developed a variational bound tailored to logistic regression. Blei and Lafferty (2007) later adapted their idea to nonconjugate topic models, and researchers have continued to improve the original bound (Khan et al., 2010; Marlin et al., 2011; Ermis and Bouchard, 2014). In other work, Braun and McAuliffe (2010) derived a variational inference algorithm for the discrete choice model, which also lies outside of the class of conditionally conjugate models. They developed a delta method to approximate the difficult-to-compute expectations. Finally, Wand et al. (2011) use auxiliary variable methods, quadrature, and mixture approximations to handle a variety of likelihood terms that fall outside of the exponential family.

More recently, researchers have generalized nonconjugate inference, seeking recipes that can be used across many models. Wang and Blei (2013) adapted Laplace approximations and the delta method to this end, improving inference in nonconjugate generalized linear models and topic models; this approach is also used by Bugbee et al. (2016) for semi-parametric regression. Knowles and Minka (2011) generalized the Jordan-Jaakkola bound in a message-passing algorithm and Wand (2014) further simplified and extended their approach. Tan and Nott (2013, 2014) applied these message-passing methods to generalized linear mixed models (and also combined them with SVI). Rohde and Wand (2015) unified many of these algorithmic developments and provided practical insights into their numerical implementations.

Finally, there has been a flurry of research on optimizing difficult variational objectives with Monte Carlo (MC) estimates of the gradient. The idea is to write the gradient of the ELBO as an expectation, compute MC estimates of it, and then use stochastic optimization with repeated MC gradients. This first appeared independently in several papers (Ji et al.,

2010; Nott et al., 2012; Paisley et al., 2012; Wingate and Weber, 2013). The newest approaches avoid any model-specific derivations, and are termed "black box" inference methods. As examples, see Kingma and Welling (2014); Rezende et al. (2014); Ranganath et al. (2014, 2016); Salimans and Knowles (2014); Titsias and Lázaro-Gredilla (2014), and Tran et al. (2016). Kucukelbir et al. (2016) leverage these ideas toward an automatic VI technique that works on any model written in the probabilistic programming system Stan (Stan Development Team, 2015). This is a step towards a derivation-free, easy-to-use VI algorithm.

5.4 Open problems

There are many open avenues for statistical research in variational inference.

We focused on optimizing $\text{KL}\left(q(\mathbf{z})||p(\mathbf{z}\,|\,\mathbf{x})\right)$ as the variational objective function. A promising avenue of research is to develop variational inference methods that optimize other measures, such as α-divergence measures. As one example, expectation propagation (Minka, 2001) is inspired by the KL divergence "in the other direction," between $p(\mathbf{z}\,|\,\mathbf{x})$ and $q(\mathbf{z})$. Other work has developed divergences based on lower bounds that are tighter than the ELBO (Barber and de van Laar, 1999; Leisink and Kappen, 2001). While alternative divergences may be difficult to optimize, they may give better approximations (Minka et al., 2005; Opper and Winther, 2005).

Though it is flexible, the mean-field family makes strong independence assumptions. These assumptions help with scalable optimization, but they limit the expressibility of the variational family. Further, they can exacerbate issues with local optima of the objective and underestimating posterior variances; see Figure 1. A second avenue of research is to develop better approximations while maintaining efficient optimization.

As we mentioned above, structured variational inference has its roots in the early days of the method (Saul and Jordan, 1996; Barber and Wiegerinck, 1999). More recently, Hoffman and Blei (2015) use generic structured variational inference in a stochastic optimization algorithm; Kucukelbir et al. (2016) and Tan and Nott (2016) take advantage of Gaussian variational families with non-diagonal covariance; Giordano et al. (2015) post-process the mean-field parameters to correct for underestimating the variance; and Ranganath et al. (2016) embed the mean-field parameters themselves in a hierarchical model to induce variational dependencies between latent variables.

Finally, the statistical properties of variational inference are not yet well understood, especially in contrast to the wealth of analysis of MCMC techniques. There has been some progress; see Section 5.2. A final open research problem is to understand variational inference as an estimator and to understand its statistical profile relative to the exact posterior.

A Bayesian Linear Regression with Automatic Relevance Determination

Consider a dataset of $\mathbf{y} = y_{1:n} \in \mathbb{R}^n$ outputs and $\mathbf{x} = x_{1:n} \in \mathbb{R}^{(n \times D)}$ D-dimensional inputs, where each $x_i \in \mathbb{R}^D$.

A linear regression model assumes a linear relationship between the inputs and the conditional mean of the output given the inputs. The latent variable $\beta \in \mathbb{R}^D$ is a vector of the regression coefficients.

Automatic relevance determination (ARD) assigns a separate prior for each regression coefficient; the idea is to automatically shrink irrelevant coefficients during inference (MacKay, 1992; Neal, 2012; Tipping, 2001; Wipf and Nagarajan, 2008). ARD works by pairing the prior precision of each regression coefficient with its own latent variable α_d. The hyper-prior on these relevance variables α encourages small values; this, in turn, selects relevant regression coefficients.

Here we present a conditionally conjugate Bayesian linear regression model with an ARD prior, based on Drugowitsch (2013). All Gaussian distributions below follow the precision parameterization.

Define a Gaussian likelihood with precision parameter τ as

$$p(\mathbf{y} \mid \beta, \tau \,;\, \mathbf{x}) = \prod_{i=1}^{n} \mathcal{N}(y_i \mid \beta^\top x_i, \tau).$$

ARD then posits the following hierarchical prior

$$p(\beta, \tau \mid \alpha) = \mathcal{N}(\beta \mid 0, \tau \operatorname{diag}(\alpha)) \operatorname{Gam}(\tau \mid a_0, b_0),$$

where α is a D-dimensional relevance variable

$$p(\alpha) = \prod_{d=1}^{D} \operatorname{Gam}(\alpha_d \mid c_0, d_0).$$

Here a_0, b_0, c_0, and d_0 are fixed hyper-parameters. The latent variable α determines the relevance of each regression coefficient.

The posterior $p(\beta, \tau, \alpha \mid \mathbf{y} \,;\, \mathbf{x})$ is not available in closed form. A simpler model that does not include α admits a closed form posterior because the normal-gamma distribution is conjugate to a normal likelihood with unknown mean and precision.

We derive a CAVI algorithm for this model. First, factorize the variational approximation as

$$q(\beta, \tau, \alpha) = q(\beta, \tau) q(\alpha).$$

Here we consider β and τ in a single factor.

Begin by applying Equation (18) to identify the optimal form of $q(\beta, \tau)$ as

$$\begin{aligned}
\log q(\beta, \tau) &= \log p(\mathbf{y} \mid \beta, \tau \,;\, \mathbf{x}) + \mathbb{E}_\alpha[\log p(\beta, \tau \mid \alpha)] + \text{const.} \\
&= \left(\frac{D}{2} + a_0 - 1 + \frac{n}{2}\right) \log \tau \\
&\quad - \frac{\tau}{2}\left(\beta^\top \left(\mathbb{E}_\alpha[\operatorname{diag} \alpha] + \sum_i x_i x_i^\top\right) \beta + \sum_i y_i^2 - 2\beta^\top \sum_i x_i y_i + 2b_0\right) \\
&\quad + \text{const.} \\
&= \log \mathcal{N}(\beta \mid \beta_*, \tau V_*^{-1}) + \log \operatorname{Gam}(\tau \mid a_*, b_*).
\end{aligned}$$

The optimal variational approximation to the regression coefficients and the precision is thus a normal-gamma with the following parameters:

$$\begin{aligned} V_*^{-1} &= \mathbb{E}_\alpha[\operatorname{diag}\alpha] + \sum_i x_i x_i^\top, \\ \beta_* &= V_* \sum_i x_i y_i, \\ a_* &= a_0 + \frac{n}{2}, \\ b_* &= b_0 + \frac{1}{2}\left(\sum_i y_i^2 - \beta_*^\top V_*^{-1} \beta_*\right). \end{aligned}$$

Next consider the optimal form of the relevance variables α. Again, apply Equation (18) to identify the optimal form of $q(\alpha) = \prod_{d=1}^D q(\alpha_d)$ as

$$\begin{aligned} \log q(\alpha_d) &= \mathbb{E}_{\beta,\tau}[\log p(\beta, \tau \mid \alpha_d)] + \log p(\alpha_d) + \text{const.} \\ &= \left(c_0 - 1 + \frac{D}{2}\right) \log \alpha_d - \alpha_d \left(d_0 + \frac{1}{2}\mathbb{E}_{\beta,\tau}[\tau \beta_d^2]\right) + \text{const.} \\ &= \log \operatorname{Gam}(\alpha_d \mid c_*, d_{*d}). \end{aligned}$$

The optimal variational approximation to the relevance variable is thus a Gamma with the following parameters:

$$\begin{aligned} c_* &= c_0 + \frac{1}{2}, \\ d_{*d} &= d_0 + \frac{1}{2}\mathbb{E}_{\beta,\tau}[\tau \beta_d^2]. \end{aligned}$$

Finally, compute the expectations as

$$\begin{aligned} \mathbb{E}_\alpha[\operatorname{diag}\alpha] &= c_* \operatorname{diag} 1/d_*, \\ \mathbb{E}_{\beta,\tau}[\tau \beta_d^2] &= \beta_{*d}^2 a_*/b_* + [V_*]_d, \end{aligned}$$

where $[\cdot]_d$ indicates the dth diagonal entry of a matrix.

Iteratively updating $a_*, b_*, c_*, d_*, V_*^{-1}$, and β_* defines CAVI for this model. These quantities also define the ELBO; Drugowitsch (2013) presents the additional algebra that computes the ELBO.

B Gaussian Mixture Model of Image Histograms

We present a multivariate (D-dimensional), diagonal covariance Gaussian mixture model (GMM). Denote a dataset of n observations as $\mathbf{x} = x_{1:n} \in \mathbb{R}^{(n \times D)}$, where each $x_i \in \mathbb{R}^D$. Assume K mixture components.

The cluster assignment latent variables are $\mathbf{z} = z_{1:n} \in \mathbb{R}^{(n \times K)}$ where each z_i is a K-indicator vector. The cluster assignments depend on the mixing vector latent variable π, which lives in a K-simplex.

The mean latent variables are $\boldsymbol{\mu} = \mu_{1:K} \in \mathbb{R}^{(K \times D)}$, where each $\mu_k \in \mathbb{R}^D$, and the precision latent variables are $\boldsymbol{\tau} = \tau_{1:K} \in \mathbb{R}^{(K \times D)}$, where each $\tau_k \in \mathbb{R}^D_{>0}$.

The joint distribution of the model factorizes as

$$p(\mathbf{x}, \mathbf{z}, \pi, \boldsymbol{\mu}, \boldsymbol{\tau}) = p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) p(\mathbf{z} \mid \pi) p(\pi) p(\boldsymbol{\mu} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau}).$$

The likelihood is Gaussian with precision parameterization

$$p(\mathbf{x} \mid \mathbf{z}, \boldsymbol{\mu}, \boldsymbol{\tau}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left(\prod_{d=1}^{D} \mathcal{N}(x_{id} \mid \mu_{kd}, \tau_{kd}) \right)^{z_{nk}}.$$

The marginal over cluster assignments is a categorical distribution,

$$p(\mathbf{z} \mid \pi) = \prod_{i=1}^{n} \prod_{k=1}^{K} \pi_k^{z_{ik}}.$$

The prior over the mixing vector is a Dirichlet distribution with fixed hyperparameters a_0,

$$p(\pi) = \mathrm{Dir}(\pi \mid a_0) = C(a_0) \prod_{k=1}^{K} \pi_k^{a_0 - 1}.$$

The prior over mean and precision parameters is a normal-gamma distribution with hyperparameters m_0, b_0, α_0, β_0,

$$\begin{aligned} p(\boldsymbol{\mu} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau}) &= \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(\mu_{kd} \mid m_0, b_0 \tau_{kd}) \times \prod_{k=1}^{K} \prod_{d=1}^{D} \mathrm{Gam}(\tau_{kd} \mid \alpha_0, \beta_0) \\ &= \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}(\mu_{kd} \mid m_0, b_0 \tau_{kd}) \, \mathrm{Gam}(\tau_{kd} \mid \alpha_0, \beta_0). \end{aligned}$$

We use the following values for the hyperparameters

$$a_0 = \frac{1}{K}, \qquad m_0 = 0.0, \qquad b_0 = 1.0, \qquad \alpha_0 = 1.0, \qquad \beta_0 = 1.0.$$

Bishop (2006, Chapter 10.2) derives a CAVI algorithm for this model.

Figure 8 presents Stan code that implements this model. Since Stan does not support discrete latent variables, we marginalize over the assignment variables.

```
data {
  int<lower=0> N; // number of data points in dataset
  int<lower=0> K; // number of mixture components
  int<lower=0> D; // dimension
  vector[D] x[N]; // observations
}

transformed data {
  vector<lower=0>[K] alpha0_vec;
  for (k in 1:K) {           // convert the scalar dirichlet prior 1/K
    alpha0_vec[k] <- 1.0/K;  // to a vector
  }
}

parameters {
  simplex[K] theta;             // mixing proportions
  vector[D] mu[K];              // locations of mixture components
  vector<lower=0>[D] sigma[K];  // standard deviations of mixture components
}

model {
  // priors
  theta ~ dirichlet(alpha0_vec);
  for (k in 1:K) {
      mu[k] ~ normal(0.0, 1.0/sigma[k]);
      sigma[k] ~ inv_gamma(1.0, 1.0);
  }

  // likelihood
  for (n in 1:N) {
    real ps[K];
    for (k in 1:K) {
      ps[k] <- log(theta[k]) + normal_log(x[n], mu[k], sigma[k]);
    }
    increment_log_prob(log_sum_exp(ps));
  }
}
```

Figure 8: Stan code for the GMM of image histograms.

C Latent Dirichlet Allocation

Probabilistic topic models are mixed-membership models of text, used to uncover the latent "topics" that run through a collection of documents. Topic models have become a popular technique for exploratory data analysis of large collections (Blei, 2012).

Latent Dirichlet allocation (LDA) (Blei et al., 2003) is a conditionally conjugate topic model (Section 4.2). It treats documents as containing multiple topics, where a topic is a distribution over words in a vocabulary.

Following the notation of Hoffman et al. (2013), let K be a specific number of topics and V the size of the vocabulary. LDA defines the following generative process:

1. For each topic in $k = 1, \ldots, K$,
   (a) draw a distribution over words $\beta_k \sim \text{Dir}_V(\eta)$.
2. For each document in $d = 1, \ldots, D$,
   (a) draw a vector of topic proportions $\theta_d \sim \text{Dir}_K(\alpha)$.
   (b) For each word in $n = 1, \ldots, N$,
      i. draw a topic assignment $z_{dn} \sim \text{Mult}(\theta_d)$, then
      ii. draw a word $w_{dn} \sim \text{Mult}(\beta_{z_{dn}})$.

Here $\eta \in \mathbb{R}_{>0}$ is a fixed parameter of the symmetric Dirichlet prior on the topics β, and $\alpha \in \mathbb{R}^K_{>0}$ are fixed parameters of the Dirichlet prior on the topic proportions for each document. Similar to the GMM example in Section 3, we encode categorical variables as indicator vectors. Thus z_{dn} is a K-vector where $z^k_{dn} = 1$ indicates the nth word in document d is assigned to the kth topic. Similarly, w_{dn} is a V-vector where $w^v_{dn} = 1$ indicates that the nth word in document d is the vth word in the vocabulary. We occasionally abuse these indicator vectors as indices—for example, if $z^k_{dn} = 1$, then $\beta_{z_{dn}}$ is the kth topic, denoted by β_k.

The posterior $p(\beta, \theta, z \mid w)$ is not available in closed form. While the topic assignments z and their proportions θ enjoy a conjugate relationship, the introduction of the topics β renders this posterior analytically intractable.

We derive a CAVI algorithm for this model, based on Hoffman et al. (2013). Posit a mean-field variational family

$$q(\beta, \theta, z) = \prod_{k=1}^{K} q(\beta_k \,;\, \lambda_k) \prod_{d=1}^{D} \left(q(\theta_d \,;\, \gamma_d) \prod_{n=1}^{N} q(z_{dn} \,;\, \phi_{dn}) \right).$$

Since LDA is a conditionally conjugate model, we can directly identify the family of each factor (Section 4.2).

Begin with the complete conditional of the topic assignment. This is a multinomial,

$$p(z_{dn} = k \mid \theta_d, \beta, w_{dn}) \propto \exp\left(\log \theta_{dk} + \log \beta_{k, w_{dn}} \right).$$

The variational approximation to the topic assignments is also a multinomial distribution, with parameters ϕ_{dn}.

Follow with the complete conditional of the topic proportions. This is a K-dimensional Dirichlet

$$p(\theta_d \mid z_d) = \text{Dir}_K \left(\alpha + \sum_{n=1}^{N} z_{dn}. \right)$$

The variational approximation to the topic proportions is also a K-dimensional Dirichlet with parameters γ_d.

End with the complete conditional of the topics. This is a V-dimensional Dirichlet

$$p(\beta_k \mid z, w) = \text{Dir}_V\left(\eta + \sum_{d=1}^{D}\sum_{n=1}^{N} z_{dn}^k w_{dn}\right).$$

In words, the vth element of the kth topic is a Dirichlet with parameter equal to the sum of the fixed scalar η and the number of times term v (denoted by w_{dn}) was assigned to topic k (denoted by z_{dn}^k). The variational approximation to the topic proportions is a V-dimensional Dirichlet with parameters λ_k.

The CAVI updates for the topic assignment and topic proportions require iterating over the following for each word within each document until convergence:

$$\begin{aligned}\phi_{dn}^k &\propto \exp\left(\mathbb{E}\left[\log\theta_{dk}\right] + \mathbb{E}\left[\log\beta_{k,w_{dn}}\right]\right) \\ &\propto \exp\left(\Psi(\gamma_{dk}) + \Psi(\lambda_{k,w_{dn}}) - \Psi\left(\sum_v \lambda_{kv}\right)\right) \qquad (59)\end{aligned}$$

$$\gamma_d = \alpha + \sum_{n=1}^{N}\phi_{dn} \qquad (60)$$

This is a direct application of Equation (47) to the complete conditionals above.

The updates for ϕ and γ depend on the variational parameters for the topics λ. The update for the topics, in turn, depends on the variational parameters for the topic proportions. That update is

$$\lambda_k = \eta + \sum_{d=1}^{D}\sum_{n=1}^{N}\phi_{dn}^k w_{dn}. \qquad (61)$$

This update only depends on the variational parameter for the topic assignments ϕ_{dn}.

Algorithm 4 presents the full CAVI algorithm for LDA. A similar computation defines the ELBO for LDA; Hoffman et al. (2013) present the additional algebra for the ELBO.

Algorithm 4: CAVI for LDA

Input: LDA model and a set of words in documents w.

Output: Variational parameters λ, γ, ϕ.

Initialize: Variational parameters λ, γ randomly.

while *the* ELBO *has not converged* **do**
  repeat
    for *each document* d **do**
      for *each word* n **do**
        Compute updates to ϕ and γ via Equations (59) and (60).
      end
    end
  until ϕ *and* γ *have converged*;
  Compute update to λ via Equation (61).
end

References

Ahmed, A., Aly, M., Gonzalez, J., Narayanamurthy, S., and Smola, A. (2012). Scalable inference in latent variable models. In *International Conference on Web Search and Data Mining*.

Airoldi, E., Blei, D., Fienberg, S., and Xing, E. (2008). Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014.

Amari, S. (1982). Differential geometry of curved exponential families-curvatures and information loss. *The Annals of Statistics*, 10(2):357–385.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007a). Gaussian process approximations of stochastic differential equations. *Workshop on Gaussian Processes in Practice*, 1:1–16.

Archambeau, C., Opper, M., Shen, Y., Cornford, D., and Shawe-Taylor, J. (2007b). Variational inference for diffusion processes. In *Neural Information Processing Systems*.

Armagan, A., Clyde, M., and Dunson, D. (2011). Generalized beta mixtures of Gaussians. In *Neural Information Processing Systems*.

Armagan, A. and Dunson, D. (2011). Sparse variational analysis of linear mixed models for large data sets. *Statistics & Probability Letters*, 81(8):1056–1062.

Barber, D. (2012). *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Barber, D. and Chiappa, S. (2006). Unified inference for variational Bayesian linear Gaussian state-space models. In *Neural Information Processing Systems*.

Barber, D. and de van Laar, P. (1999). Variational cumulant expansions for intractable distributions. *Journal of Artificial Intelligence Research*, pages 435–455.

Barber, D. and Wiegerinck, W. (1999). Tractable variational structures for approximating graphical models. In *Neural Information Processing Systems*.

Beal, M. and Ghahramani, Z. (2003). The variational Bayesian EM algorithm for incomplete data: With application to scoring graphical model structures. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics 7*. Oxford University Press.

Bernardo, J. and Smith, A. (1994). *Bayesian Theory*. John Wiley & Sons Ltd., Chichester.

Bickel, P., Choi, D., Chang, X., Zhang, H., et al. (2013). Asymptotic normality of maximum likelihood and its variational approximation for stochastic blockmodels. *The Annals of Statistics*, 41(4):1922–1943.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Bishop, C., Lawrence, N., Jaakkola, T., and Jordan, M. I. (1998). Approximating posterior distributions in belief networks using mixtures. In *Neural Information Processing Systems*.

Bishop, C. and Winn, J. (2000). Non-linear Bayesian image modelling. In *European Conference on Computer Vision*.

Blei, D. (2012). Probabilistic topic models. *Communications of the ACM*, 55(4):77–84.

Blei, D. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Journal of Bayesian Analysis*, 1(1):121–144.

Blei, D. and Lafferty, J. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, 1(1):17–35.

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

Braun, M. and McAuliffe, J. (2010). Variational inference for large-scale models of discrete choice. *Journal of the American Statistical Association*, 105(489):324–335.

Brown, L. (1986). *Fundamentals of Statistical Exponential Families*. Institute of Mathematical Statistics, Hayward, CA.

Bugbee, B., Breidt, F., and van der Woerd, M. (2016). Laplace variational approximation for semiparametric regression in the presence of heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 25:225–245.

Carbonetto, P. and Stephens, M. (2012). Scalable variational inference for Bayesian variable selection in regression, and its accuracy in genetic association studies. *Bayesian Analysis*, 7(1):73–108.

Celisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electronic Journal of Statistics*, 6:1847–1899.

Chan, A. and Vasconcelos, N. (2009). Layered dynamic textures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1862–1879.

Cohen, S. and Smith, N. (2010). Covariance in unsupervised learning of probabilistic grammars. *The Journal of Machine Learning Research*, 11:3017–3051.

Cummins, M. and Newman, P. (2008). FAB-MAP: Probabilistic localization and mapping in the space of appearance. *The International Journal of Robotics Research*, 27(6):647–665.

Damianou, A., Titsias, M., and Lawrence, N. (2011). Variational Gaussian process dynamical systems. In *Neural Information Processing Systems*.

Daunizeau, J., Adam, V., and Rigoux, L. (2014). VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput Biol*, 10(1):e1003441.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Deng, L. (2004). Switching dynamic system models for speech articulation and acoustics. In *Mathematical Foundations of Speech and Language Processing*, pages 115–133. Springer.

Diaconis, P., Ylvisaker, D., et al. (1979). Conjugate priors for exponential families. *The Annals of Statistics*, 7(2):269–281.

Drugowitsch, J. (2013). Variational Bayesian inference for linear and logistic regression. *arXiv preprint arXiv:1310.5438*.

Du, L., Lu, R., Carin, L., and Dunson, D. (2009). A Bayesian model for simultaneous image clustering, annotation and object segmentation. In *Neural Information Processing Systems*.

Ermis, B. and Bouchard, G. (2014). Iterative splits of quadratic bounds for scalable binary tensor factorization. In *Uncertainty in Artificial Intelligence*.

Erosheva, E. A., Fienberg, S. E., and Joutard, C. (2007). Describing disability through individual-level mixture models for multivariate binary data. *The Annals of Applied Statistics*, 1(2):346–384.

Flandin, G. and Penny, W. (2007). Bayesian fMRI data analysis with sparse spatial basis function priors. *NeuroImage*, 34(3):1108–1125.

Foti, N., Xu, J., Laird, D., and Fox, E. (2014). Stochastic variational inference for hidden Markov models. In *Neural Information Processing Systems*.

Furmston, T. and Barber, D. (2010). Variational methods for reinforcement learning. In *Artificial Intelligence and Statistics*.

Gelfand, A. and Smith, A. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6:721–741.

Gershman, S. J., Blei, D. M., Norman, K. A., and Sederberg, P. B. (2014). Decomposing spatiotemporal brain patterns into topographic latent sources. *NeuroImage*, 98:91–102.

Ghahramani, Z. and Jordan, M. I. (1997). Factorial hidden Markov models. *Machine Learning*, 29(2-3):245–273.

Giordano, R. J., Broderick, T., and Jordan, M. I. (2015). Linear response methods for accurate covariance estimates from mean field variational Bayes. In *Neural Information Processing Systems*.

Grimmer, J. (2011). An introduction to Bayesian inference via variational approximations. *Political Analysis*, 19(1):32–47.

Hall, P., Ormerod, J., and Wand, M. (2011a). Theory of Gaussian variational approximation for a Poisson mixed model. *Statistica Sinica*, 21:369–389.

Hall, P., Pham, T., Wand, M., and Wang, S. (2011b). Asymptotic normality and valid inference for Gaussian variational approximation. *Annals of Statistics*, 39(5):2502–2532.

Harrison, L. and Green, G. (2010). A Bayesian spatiotemporal model for very large data sets. *Neuroimage*, 50(3):1126–1141.

Hastings, W. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Hensman, J., Fusi, N., and Lawrence, N. (2014). Gaussian processes for big data. In *Uncertainty in Artificial Intelligence*.

Hensman, J., Rattray, M., and Lawrence, N. (2012). Fast variational inference in the conjugate exponential family. In *Neural Information Processing Systems*.

Hinton, G. and Van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *Computational Learning Theory*.

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

Hoffman, M. D. and Blei, D. M. (2015). Structured stochastic variational inference. In *Artificial Intelligence and Statistics*.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623.

Honkela, A., Tornio, M., Raiko, T., and Karhunen, J. (2008). Natural conjugate gradient in variational inference. In *Neural Information Processing*, pages 305–314. Springer.

Jaakkola, T. and Jordan, M. I. (1996). Computing upper and lower bounds on likelihoods in intractable networks. In *Uncertainty in Artificial Intelligence*.

Jaakkola, T. and Jordan, M. I. (1997). A variational approach to Bayesian logistic regression models and their extensions. In *Artificial Intelligence and Statistics*.

Jaakkola, T. and Jordan, M. I. (2000). Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37.

Ji, C., Shen, H., and West, M. (2010). Bounded approximations for marginal likelihoods. Technical report, Duke University.

Johnson, M. and Willsky, A. (2014). Stochastic variational inference for Bayesian time series models. In *International Conference on Machine Learning*.

Jojic, N. and Frey, B. (2001). Learning flexible sprites in video layers. In *Computer Vision and Pattern Recognition*.

Jojic, V., Jojic, N., Meek, C., Geiger, D., Siepel, A., Haussler, D., and Heckerman, D. (2004). Efficient approximations for learning phylogenetic HMM models from data. *Bioinformatics*, 20:161–168.

Jordan, M. I., Ghahramani, Z., Jaakkola, T., and Saul, L. (1999). Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Khan, M. E., Bouchard, G., Murphy, K. P., and Marlin, B. M. (2010). Variational bounds for mixed-data factor analysis. In *Neural Information Processing Systems*.

Kiebel, S., Daunizeau, J., Phillips, C., and Friston, K. (2008). Variational Bayesian inversion of the equivalent current dipole model in EEG/MEG. *NeuroImage*, 39(2):728–741.

Kingma, D. and Welling, M. (2014). Auto-encoding variational Bayes. In *International Conference on Learning Representations*.

Knowles, D. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. In *Neural Information Processing Systems*.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in Stan. In *Neural Information Processing Systems*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2016). Automatic differentiation variational inference. *arXiv preprint arXiv:1603.00788*.

Kullback, S. and Leibler, R. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Kurihara, K. and Sato, T. (2006). Variational Bayesian grammar induction for natural language. In *Grammatical Inference: Algorithms and Applications*, pages 84–96. Springer.

Kushner, H. and Yin, G. (1997). *Stochastic Approximation Algorithms and Applications*. Springer New York.

Lashkari, D., Sridharan, R., Vul, E., Hsieh, P., Kanwisher, N., and Golland, P. (2012). Search for patterns of functional specificity in the brain: A nonparametric hierarchical Bayesian model for group fMRI data. *Neuroimage*, 59(2):1348–1368.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B*, pages 157–224.

Le Cun, Y. and Bottou, L. (2004). Large scale online learning. In *Neural Information Processing Systems*.

Leisink, M. and Kappen, H. (2001). A tighter bound for graphical models. *Neural Computation*, 13(9):2149–2171.

Liang, P., Jordan, M. I., and Klein, D. (2009). Probabilistic grammars and hierarchical Dirichlet processes. In O'Hagan, T. and West, M., editors, *The Handbook of Applied Bayesian Analysis*. New York: Oxford Univ. Press.

Liang, P., Petrov, S., Klein, D., and Jordan, M. I. (2007). The infinite PCFG using hierarchical Dirichlet processes. In *Empirical Methods in Natural Language Processing*.

Likas, A. and Galatsanos, N. (2004). A variational approach for Bayesian blind image deconvolution. *IEEE Transactions on Signal Processing*, 52(8):2222–2233.

Logsdon, B., Hoffman, G., and Mezey, J. (2010). A variational Bayes algorithm for fast and accurate multiple locus genome-wide association analysis. *BMC Bioinformatics*, 11(1):58.

MacKay, D. J. (1992). Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D. J. (1997). Ensemble learning for hidden Markov models. Unpublished manuscript from http://www.inference.eng.cam.ac.uk/mackay/ensemblePaper.pdf.

Manning, J. R., Ranganath, R., Norman, K. A., and Blei, D. M. (2014). Topographic factor analysis: a Bayesian model for inferring brain networks from neural data. *PloS one*, 9(5):e94914.

Marlin, B. M., Khan, M. E., and Murphy, K. P. (2011). Piecewise bounds for estimating Bernoulli-logistic latent Gaussian models. In *International Conference on Machine Learning*.

McGrory, C. A. and Titterington, D. M. (2007). Variational approximations in Bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis*, 51(11):5352–5367.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, M., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092.

Minka, T. (2005). Divergence measures and message passing. Technical Report TR-2005-173, Microsoft Research.

Minka, T. et al. (2005). Divergence measures and message passing. Technical report, Technical report, Microsoft Research.

Minka, T., Winn, J., Guiver, J., Webster, S., Zaykov, Y., Yangel, B., Spengler, A., and Bronskill, J. (2014). Infer.NET 2.6.

Minka, T. P. (2001). Expectation propagation for approximate Bayesian inference. In *Uncertainty in Artificial Intelligence*.

Naseem, T., Chen, H., Barzilay, R., and Johnson, M. (2010). Using universal linguistic knowledge to guide grammar induction. In *Empirical Methods in Natural Language Processing*.

Nathoo, F., Babul, A., Moiseev, A., Virji-Babul, N., and Beg, M. (2014). A variational Bayes spatiotemporal model for electromagnetic brain mapping. *Biometrics*, 70(1):132–143.

Neal, R. and Hinton, G. (1999). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. MIT Press.

Neal, R. M. (2012). *Bayesian Learning for Neural Networks*. Springer Science & Business Media.

Neville, S., Ormerod, J., and Wand, M. (2014). Mean field variational Bayes for continuous sparse signal shrinkage: pitfalls and remedies. *Electronic Journal of Statistics*, 8(1):1113–1151.

Nott, D. J., Tan, S. L., Villani, M., and Kohn, R. (2012). Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 21(3):797–820.

Opper, M. and Winther, O. (2005). Expectation consistent approximate inference. *The Journal of Machine Learning Research*, 6:2177–2204.

Ormerod, J., You, C., and Muller, S. (2014). A variational Bayes approach to variable selection. Unpublished manuscript from http://www.maths.usyd.edu.au/u/jormerod/JTOpapers/VariableSelectionFinal.pdf.

Paisley, J., Blei, D., and Jordan, M. I. (2012). Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Penny, W., Kiebel, S., and Friston, K. (2003). Variational Bayesian inference for fMRI time series. *NeuroImage*, 19(3):727–741.

Penny, W., Trujillo-Barreto, N., and Friston, K. (2005). Bayesian fMRI time series analysis with spatial priors. *Neuroimage*, 24:350–362.

Peterson, C. and Anderson, J. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, 1(5):995–1019.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Ramos, F., Upcroft, B., Kumar, S., and Durrant-Whyte, H. (2012). A Bayesian approach for place recognition. *Robotics and Autonomous Systems*, 60(4):487–497.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*.

Ranganath, R., Tran, D., and Blei, D. (2016). Hierarchical variational models. In *International Conference on Machine Learning*.

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

Reyes-Gomez, M., Ellis, D., and Jojic, N. (2004). Multiband audio modeling for single-channel acoustic source separation. In *Acoustics, Speech, and Signal Processing*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robert, C. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag, New York, NY.

Roberts, S., Guilford, T., Rezek, I., and Biro, D. (2004). Positional entropy during pigeon homing I: Application of Bayesian latent state modelling. *Journal of Theoretical Biology*, 227:39–50.

Roberts, S. and Penny, W. (2002). Variational Bayes for generalized autoregressive models. *IEEE Transactions on Signal Processing*, 50(9):2245–2257.

Rohde, D. and Wand, M. (2015). Semiparametric mean field variational Bayes: General principles and numerical issues. Unpublished manuscript from http://mattwand.utsacademics.info/RohdeWand.pdf.

Salimans, T. and Knowles, D. (2014). On using control variates with stochastic approximation for variational Bayes. *arXiv preprint arXiv:1401.1022*.

Sanguinetti, G., Lawrence, N., and Rattray, M. (2006). Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781.

Sato, M. (2001). Online model selection based on the variational Bayes. *Neural Computation*, 13(7):1649–1681.

Sato, M., Yoshioka, T., Kajihara, S., Toyama, K., Goda, N., Doya, K., and Kawato, M. (2004). Hierarchical Bayesian estimation for MEG inverse problem. *NeuroImage*, 23(3):806–826.

Saul, L. and Jordan, M. I. (1996). Exploiting tractable substructures in intractable networks. In *Neural Information Processing Systems*.

Saul, L. K., Jaakkola, T., and Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4(1):61–76.

Spall, J. (2003). *Introduction to Stochastic Search and Optimization: Estimation, Simulation, and Control*. John Wiley and Sons.

Stan Development Team (2015). *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*.

Stegle, O., Parts, L., Durbin, R., and Winn, J. (2010). A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eqtl studies. *PLoS Comput Biol*, 6(5):e1000770.

Sudderth, E. B. and Jordan, M. I. (2009). Shared segmentation of natural scenes using dependent Pitman-Yor processes. In *Neural Information Processing Systems*.

Sung, J., Ghahramani, Z., and Bang, Y. (2008). Latent-space variational Bayes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(12):2236–2242.

Sykacek, P., Roberts, S., and Stokes, M. (2004). Adaptive BCI based on variational Bayesian Kalman filtering: An empirical evaluation. *IEEE Transactions on Biomedical Engineering*, 51(5):719–727.

Tan, L. and Nott, D. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188.

Tan, L. and Nott, D. (2014). A stochastic variational framework for fitting and diagnosing generalized linear mixed models. *Bayesian Analysis*, 9(4):963–1004.

Tan, L. and Nott, D. (2016). Gaussian variational approximation with sparse precision matrix. *arXiv:1605.05622*.

Tipping, M. and Lawrence, N. (2005). Variational inference for Student-t models: Robust Bayesian interpolation and generalised component analysis. *Neurocomputing*, 69(1):123–141.

Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1(Jun):211–244.

Titsias, M. and Lawrence, N. (2010). Bayesian Gaussian process latent variable model. In *Artificial Intelligence and Statistics*.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*.

Tran, D., Ranganath, R., and Blei, D. M. (2016). The variational Gaussian process. In *International Conference on Learning Representations*.

Ueda, N. and Ghahramani, Z. (2002). Bayesian model search for mixture models based on optimizing variational bounds. *Neural Networks*, 15(10):1223–1241.

Van Den Broek, B., Wiegerinck, W., and Kappen, B. (2008). Graphical model inference in optimal control of stochastic multi-agent systems. *Journal of Artificial Intelligence Research*, 32:95–122.

Vermaak, J., Lawrence, N. D., and Pérez, P. (2003). Variational inference for visual tracking. In *Computer Vision and Pattern Recognition*.

Villegas, M., Paredes, R., and Thomee, B. (2013). Overview of the ImageCLEF 2013 Scalable Concept Image Annotation Subtask. In *CLEF Evaluation Labs and Workshop*.

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

Wand, M. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, 15:1351–1369.

Wand, M., Ormerod, J., Padoan, S., and Fuhrwirth, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6:847–900.

Wang, B. and Titterington, D. (2005). Inadequacy of interval estimates corresponding to variational Bayesian approximations. In *Artificial Intelligence and Statistics*.

Wang, B. and Titterington, D. (2006). Convergence properties of a general algorithm for calculating variational Bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1:625–650.

Wang, C. and Blei, D. (2013). Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:1005–1031.

Wang, C. and Blei, D. (2015). A general method for robust Bayesian modeling. *arXiv preprint arXiv:1510.05078*.

Wang, P. and Blunsom, P. (2013). Collapsed variational Bayesian inference for hidden Markov models. In *Artificial Intelligence and Statistics*.

Wang, Y. and Mori, G. (2009). Human action recognition by semilatent topic models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(10):1762–1774.

Waterhouse, S., MacKay, D., and Robinson, T. (1996). Bayesian methods for mixtures of experts. *Neural Information Processing Systems*.

Welling, M. and Teh, Y. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *International Conference on Machine Learning*.

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

Wiggins, C. and Hofman, J. (2008). Bayesian approach to network modularity. *Physical Review Letters*, 100(25).

Wingate, D. and Weber, T. (2013). Automated variational inference in probabilistic programming. *arXiv preprint arXiv:1301.1299*.

Winn, J. and Bishop, C. (2005). Variational message passing. *Journal of Machine Learning Research*, 6:661–694.

Wipf, D. and Nagarajan, S. (2009). A unified Bayesian framework for MEG/EEG source imaging. *NeuroImage*, 44(3):947–966.

Wipf, D. P. and Nagarajan, S. S. (2008). A new view of automatic relevance determination. In *Neural Information Processing Systems*.

Woolrich, M., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. (2004). Multilevel linear modeling for fMRI group analysis using Bayesian inference. *Neuroimage*, 21:1732–1747.

Xing, E., Wu, W., Jordan, M. I., and Karp, R. (2004). Logos: A modular Bayesian model for de novo motif detection. *Journal of Bioinformatics and Computational Biology*, 2(01):127–154.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *Neural Information Processing Systems*.

Yogatama, D., Wang, C., Routledge, B., Smith, N. A., and Xing, E. (2014). Dynamic language models for streaming text. *Transactions of the Association for Computational Linguistics*, 2:181–192.

You, C., Ormerod, J., and Muller, S. (2014). On variational Bayes estimation and variational information criteria for linear regression models. *Australian & New Zealand Journal of Statistics*, 56(1):73–87.

Yu, T. and Wu, Y. (2005). Decentralized multiple target tracking using netted collaborative autonomous trackers. In *Computer Vision and Pattern Recognition*.

Zumer, J., Attias, H., Sekihara, K., and Nagarajan, S. (2007). A probabilistic algorithm integrating source localization and noise suppression for MEG and EEG data. *NeuroImage*, 37(1):102–115.