# Explaining Variational Approximations

J. T. Ormerod and M. P. Wand

Variational approximations facilitate approximate inference for the parameters in complex statistical models and provide fast, deterministic alternatives to Monte Carlo methods. However, much of the contemporary literature on variational approximations is in Computer Science rather than Statistics, and uses terminology, notation, and examples from the former field. In this article we explain variational approximation in statistical terms. In particular, we illustrate the ideas of variational approximation using examples that are familiar to statisticians.

KEY WORDS: Bayesian inference; Bayesian networks; Directed acyclic graphs; Generalized linear mixed models; Kullback–Leibler divergence; Linear mixed models.

## 1. INTRODUCTION

*Variational approximations* is a body of deterministic techniques for making approximate inference for parameters in complex statistical models. It is now part of mainstream Computer Science methodology, where it enjoys use in elaborate problems such as speech recognition, document retrieval, and genetic linkage analysis (Jordan 2004). Summaries of contemporary variational approximations can be found in the works of Jordan et al. (1999), Jordan (2004), Titterington (2004), and Bishop (2006, chap. 10). In 2008, a variational approximation-based software package named Infer.NET (Minka et al. 2009) emerged with claims of being able to handle a wide variety of statistical problems.

The name 'variational approximations' has its roots in the mathematical topic known as *variational calculus*. Variational calculus is concerned with the problem of optimizing a functional over a class of functions on which that functional depends. Approximate solutions arise when the class of functions is restricted in some way—usually to enhance tractability.

Despite their statistical overtones, variational approximations are not widely known within the statistical community. In particular, they are overshadowed by Monte Carlo methods, especially Markov chain Monte Carlo (MCMC), for performing approximate inference, as well as Laplace approximation methods. Variational approximations are a much faster alternative to MCMC, especially for large models, and are a richer class of methods than the Laplace variety. They are, however, limited in their approximation accuracy—as opposed to MCMC which can be made arbitrarily accurate through increases in the Monte Carlo sample sizes. In the interest of brevity, we will not discuss the *quality* of variational approximations in any detail. Jordan (2004) and Titterington (2004) pointed to some relevant literature on variational approximation accuracy.

In the statistics literature, variational approximations are beginning to have a presence. Examples include the articles by Teschendorff et al. (2005), McGrory and Titterington (2007), and McGrory et al. (2009) on new variational approximation methodology for particular applications, and Hall, Humphreys, and Titterington (2002) and Wang and Titterington (2006) on the statistical properties of estimators obtained via variational approximation.

In this article we explain variational approximation in terms that are familiar to a statistical readership. Most of our exposition involves working through several illustrative examples, starting with what is perhaps the most basic: inference from a Normal random sample. Other contexts that are seen to benefit from variational approximation include Bayesian generalized linear models, Bayesian linear mixed models, and non-Bayesian generalized linear mixed models. It is anticipated that a statistically literate reader who works through all of the examples will have gained a good understanding of variational approximations.

Variational approximations can be useful for both likelihood-based and Bayesian inference. However, their use in the literature is greater for Bayesian inference where intractable calculus problems abound. Hence, most of our description of variational approximations is for Bayesian inference. It is also worth noting that situations in which variational approximations are useful closely correspond to situations where MCMC is useful.

Section 2 explains the most common variant of variational approximation, which we call the *density transform* approach. A different type, the *tangent transform* approach, is explained in Section 3. Sections 2 and 3 focus exclusively on Bayesian inference. In Section 4 we point out that the same ideas transfer to frequentist contexts. Some concluding remarks are made in Section 5.

### 1.1 Definitions

Integrals without limits or subscripts are assumed to be over the entire space of the integrand argument. If $\mathcal{P}$ is a logical condition, then $I(\mathcal{P}) = 1$ if $\mathcal{P}$ is true and $I(\mathcal{P}) = 0$ if $\mathcal{P}$ is false. We use $\Phi$ and $\phi$ to denote the standard normal distribution function and density function, respectively. The Gamma function, denoted by $\Gamma$, is given by $\Gamma(x) = \int_0^\infty u^{x-1} e^{-u}\, du$

J. T. Ormerod is Lecturer in Statistics, School of Mathematics and Statistics, University of Sydney, Sydney 2006, Australia. M. P. Wand is Research Professor in Statistics, Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia (E-mail: *mwand@uow.edu.au*). The authors are grateful to the editor, an associate editor, and two referees for their suggestions for improvement. They also thank Christel Faes, Sarah Neville, Agostino Nobile, Simone Padoan, Doug Simpson, Mike Titterington, and Shen Wang for helpful comments. This research was partially supported by Australian Research Council Discovery project DP0877055.

© 2010 American Statistical Association DOI: 10.1198/tast.2010.09058

and the Digamma function, denoted by $\psi$, is given by $\psi(x) = \frac{d}{dx}\log\Gamma(x)$.

Column vectors with entries consisting of subscripted variables are denoted by a boldfaced version of the letter for that variable. Round brackets will be used to denote the entries of column vectors. For example, $\mathbf{x} = (x_1, \ldots, x_n)$ denotes an $n \times 1$ vector with entries $x_1, \ldots, x_n$. Scalar functions applied to vectors are evaluated element-wise. For example,

$$\exp(a_1, a_2, a_3) \equiv (\exp(a_1), \exp(a_2), \exp(a_3)).$$

Similarly, $(a_1, a_2, a_3)^{(b_1, b_2, b_3)} \equiv (a_1^{b_1}, a_2^{b_2}, a_3^{b_3})$. The element-wise product of two matrices $\mathbf{A}$ and $\mathbf{B}$ is denoted by $\mathbf{A} \odot \mathbf{B}$. We use $\mathbf{1}_d$ to denote the $d \times 1$ column vector with all entries equal to 1. The norm of a column vector $\mathbf{v}$, defined to be $\sqrt{\mathbf{v}^T\mathbf{v}}$, is denoted by $\|\mathbf{v}\|$. For a $d \times 1$ vector $\mathbf{a}$, we let $\text{diag}(\mathbf{a})$ denote the $d \times d$ diagonal matrix containing the entries of $\mathbf{a}$ along the main diagonal. For a $d \times d$ square matrix $\mathbf{A}$, we let $\text{diagonal}(\mathbf{A})$ denote the $d \times 1$ vector containing the diagonal entries of $\mathbf{A}$. For square matrices $\mathbf{A}_1, \ldots, \mathbf{A}_r$, we let $\text{blockdiag}(\mathbf{A}_1, \ldots, \mathbf{A}_r)$ denote the block diagonal matrix, with $i$th block equal to $\mathbf{A}_i$.

The density function of a random vector $\mathbf{u}$ is denoted by $p(\mathbf{u})$. The conditional density of $\mathbf{u}$ given $\mathbf{v}$ is denoted by $p(\mathbf{u}|\mathbf{v})$. The covariance matrix of $\mathbf{u}$ is denoted by $\text{Cov}(\mathbf{u})$. A $d \times 1$ random vector $\mathbf{x}$ has a Multivariate Normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, denoted by $\mathbf{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, if its density function is

$$p(\mathbf{x}) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}.$$

A random variable $x$ has an Inverse Gamma distribution with parameters $A, B > 0$, denoted by $x \sim \text{IG}(A, B)$, if its density function is $p(x) = B^A\Gamma(A)^{-1}x^{-A-1}e^{-B/x}, x > 0$. A random vector $\mathbf{x} = (x_1, \ldots, x_K)$ has a Dirichlet distribution with parameter vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where each $\alpha_k > 0$, if its density function is

$$p(\mathbf{x}) = \begin{cases} \left\{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\Big/\prod_{k=1}^{K}\Gamma(\alpha_k)\right\}\prod_{k=1}^{K}x_k^{\alpha_k-1}, \\ \qquad \text{if } \sum_{k=1}^{K}x_k = 1 \\ 0, \quad \text{otherwise.}\end{cases}$$

We write $\mathbf{x} \sim \text{Dirichlet}(\boldsymbol{\alpha})$. If $y_i$ has distribution $D_i$ for each $1 \le i \le n$, and the $y_i$ are independent, then we write $y_i \overset{\text{ind.}}{\sim} D_i$.

It is helpful, although not necessary, to work with *directed acyclic graph (DAG)* depictions of Bayesian statistical models. One reason is the localness of the calculations that arise from the Markov blanket result given in Section 2.2.1. The nodes of the DAG correspond to random variables or random vectors in the Bayesian model, and the directed edges convey conditional independence. Because of this connection with Bayesian (hierarchical) models, DAGs with random nodes are known as *Bayesian networks* in the Computer Science literature. Figure 1 provides DAGs corresponding to the Bayesian Poisson mixed model:

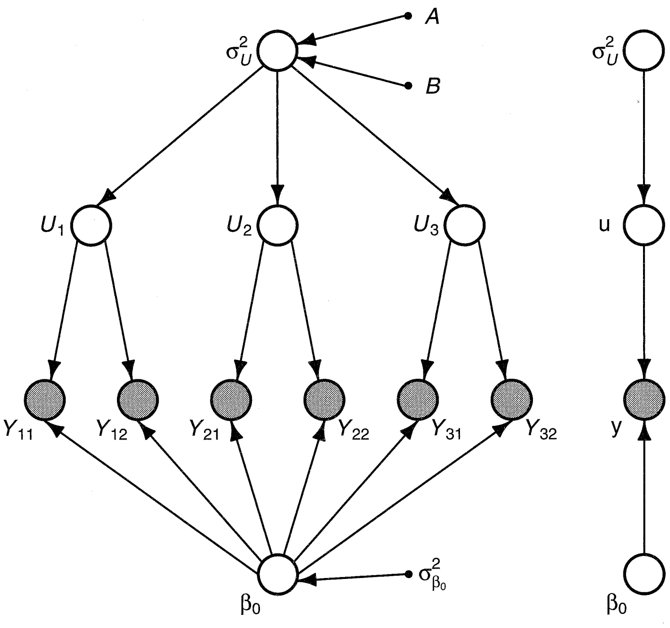

Figure 1. DAGs corresponding to the Bayesian Poisson regression model (1). *Left*: the large nodes correspond to scalar random variables in the model. The smaller nodes correspond to constants and the observed data are shaded. *Right*: abbreviated DAG for the same model. The constants are suppressed and the nodes $\mathbf{u}$ and $\mathbf{y}$ correspond to random vectors containing the $U_i$ and $y_{ij}$, respectively.

$$\begin{aligned} Y_{ij}|U_i &\overset{\text{ind.}}{\sim} \text{Poisson}(e^{\beta_0+U_i}), \qquad i = 1,2,3;\ j = 1,2, \\ U_i|\sigma_U^2 &\overset{\text{ind.}}{\sim} N(0, \sigma_U^2), \qquad \beta_0 \sim N(0, \sigma^2_{\beta_0}), \\ \sigma_U^2 &\sim \text{IG}(A, B) \quad \text{for constants } \sigma^2_{\beta_0}, A, B > 0. \end{aligned} \tag{1}$$

The DAG on the left side of Figure 1 has a separate node for each scalar random variable and each constant. On the right side the constant nodes are suppressed and two of the nodes correspond to the random vectors $\mathbf{u} \equiv (U_1, U_2, U_3)$ and $\mathbf{y} = (Y_{11}, \ldots, Y_{32})$.

## 2. DENSITY TRANSFORM APPROACH

The *density transform* approach to variational approximation involves approximation of posterior densities by other densities for which inference is more tractable. The approximations are guided by the notion of *Kullback–Leibler divergence*, which we now explain.

### 2.1 Kullback–Leibler Divergence

Consider a generic Bayesian model with parameter vector $\boldsymbol{\theta} \in \Theta$ and observed data vector $\mathbf{y}$. Bayesian inference is based on the posterior density function

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})}.$$

The denominator $p(\mathbf{y})$ is known as the *marginal likelihood* (or *model evidence* in the Computer Science literature) and forms the basis of model comparison via Bayes factors (e.g., Kass and Raftery 1995). It should be noted that, in this Bayesian context, $p(\mathbf{y})$ is not a likelihood function in the usual sense. Throughout this section we assume that $\mathbf{y}$ and $\boldsymbol{\theta}$ are continuous random vectors. The discrete case has a similar treatment, but with summations rather than integrals.

Let $q$ be an arbitrary density function over $\Theta$. Then the logarithm of the marginal likelihood satisfies

$$\begin{aligned}\log p(\mathbf{y}) &= \log p(\mathbf{y}) \int q(\boldsymbol{\theta})\, d\boldsymbol{\theta} = \int q(\boldsymbol{\theta}) \log p(\mathbf{y})\, d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{y},\boldsymbol{\theta})/q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})/q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta} \\ &= \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta} \\ &\quad + \int q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right\} d\boldsymbol{\theta} \\ &\geq \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta}. \qquad (2)\end{aligned}$$

The inequality arises from the fact that

$$\int q(\boldsymbol{\theta}) \log\left\{\frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\right\} d\boldsymbol{\theta} \geq 0$$

for all densities $q$, with equality if and only if

$$q(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{y}) \text{ almost everywhere} \qquad (3)$$

(Kullback and Leibler 1951). The integral in (3) is known as the *Kullback–Leibler divergence* (also known as *Kullback–Leibler distance*) between $q$ and $p(\cdot|\mathbf{y})$. From (2), it follows immediately that

$$p(\mathbf{y}) \geq \underline{p}(\mathbf{y}; q),$$

where the $q$-dependent lower bound on the marginal likelihood is given by

$$\underline{p}(\mathbf{y}; q) \equiv \exp \int q(\boldsymbol{\theta}) \log\left\{\frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\right\} d\boldsymbol{\theta}. \qquad (4)$$

Note that the lower bound $\underline{p}(\mathbf{y}; q)$ can also be derived more directly using Jensen's inequality, but the above derivation has the advantage of quantifying the gap between $p(\mathbf{y})$ and $\underline{p}(\mathbf{y}; q)$.

The essence of the density transform variational approach is approximation of the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$ by a $q(\boldsymbol{\theta})$ for which $\underline{p}(\mathbf{y}; q)$ is more tractable than $p(\mathbf{y})$. Tractability is achieved by restricting $q$ to a more manageable class of densities, and then maximizing $\underline{p}(\mathbf{y}; q)$ over that class. According to (2), maximization of $\underline{p}(\mathbf{y}; q)$ is equivalent to minimization of the Kullback–Leibler divergence between $q$ and $p(\cdot|\mathbf{y})$.

The most common restrictions for the $q$ density are:

(a) $q(\boldsymbol{\theta})$ factorizes into $\prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i)$, for some partition $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ of $\boldsymbol{\theta}$.

(b) $q$ is a member of a parametric family of density functions.

In the case of (a), note that the product form is the *only* assumption being made. Hence (a) represents a type of *nonparametric* restriction. Restriction (a) is also known as *mean field* approximation and has its roots in Physics (e.g., Parisi 1988). The term *variational Bayes* has become commonplace for approximate Bayesian inference under product density restrictions.

Depending on the Bayesian model at hand, both restrictions can have minor or major impacts on the resulting inference. For example, if $p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2|\mathbf{y})$ is such that $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ have a high degree of dependence, then the restriction $q(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = q_1(\boldsymbol{\theta}_1)q_2(\boldsymbol{\theta}_2)$ will lead to a degradation in the resulting inference. Conversely, if the posterior dependence between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ is weak, then the product density restriction could lead to very accurate approximate inference. Further discussion on this topic, including references, may be found in section 3.2 of the article by Titterington (2004).

## 2.2 Product Density Transforms

Restriction of $q$ to a subclass of product densities gives rise to explicit solutions for each product component in terms of the others. These, in turn, lead to an iterative scheme for obtaining the simultaneous solutions. The solutions rely on the following result, which we call Result 1. Note that Result 1 follows immediately from (2) and (3) above. However, it is useful to present the result for general random vectors.

*Result 1.* Let $\mathbf{u}$ and $\mathbf{v}$ be two continuous random vectors with joint density function $p(\mathbf{u}, \mathbf{v})$. The maximum value of

$$\int q(\mathbf{u}) \log\left\{\frac{p(\mathbf{u},\mathbf{v})}{q(\mathbf{u})}\right\} d\mathbf{u}$$

over all density functions $q$ is attained by $q^*(\mathbf{u}) = p(\mathbf{u}|\mathbf{v})$.

Return now to the Bayesian model setting of Section 2.1 and suppose that $q$ is subject to the product restriction (a). Then

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) &= \int \prod_{i=1}^{M} q_i(\boldsymbol{\theta}_i) \left\{ \log p(\mathbf{y}, \boldsymbol{\theta}) \right. \\ &\qquad \left. - \sum_{i=1}^{M} \log q_i(\boldsymbol{\theta}_i) \right\} d\boldsymbol{\theta}_1 \cdots d\boldsymbol{\theta}_M \\ &= \int q_1(\boldsymbol{\theta}_1) \left\{ \int \log p(\mathbf{y}, \boldsymbol{\theta}) q_2(\boldsymbol{\theta}_2) \cdots \right. \\ &\qquad \left. \times q_M(\boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_2 \cdots d\boldsymbol{\theta}_M \right\} d\boldsymbol{\theta}_1 \\ &\qquad - \int q_1(\boldsymbol{\theta}_1) \log q_1(\boldsymbol{\theta}_1)\, d\boldsymbol{\theta}_1 \\ &\qquad + \text{terms not involving } q_1.\end{aligned}$$

Define the new joint density function $\widetilde{p}(\mathbf{y}, \boldsymbol{\theta}_1)$ by

$$\begin{aligned}\widetilde{p}(\mathbf{y}, \boldsymbol{\theta}_1) &\equiv \exp \int \log p(\mathbf{y}, \boldsymbol{\theta}) q_2(\boldsymbol{\theta}_2) \cdots q_M(\boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_2 \cdots d\boldsymbol{\theta}_M \\ &\qquad \Big/ \int \int \left\{ \exp \int \log p(\mathbf{y}, \boldsymbol{\theta}) q_2(\boldsymbol{\theta}_2) \cdots \right. \\ &\qquad \left. \times q_M(\boldsymbol{\theta}_M)\, d\boldsymbol{\theta}_2 \cdots d\boldsymbol{\theta}_M \right\} d\boldsymbol{\theta}_1\, d\mathbf{y}.\end{aligned}$$

**Algorithm 1** Iterative scheme for obtaining the optimal densities under product density restriction (a). The updates are based on the solutions given at (5).

```
Initialize: $q_2^*(\boldsymbol{\theta}_2),\dots,q_M^*(\boldsymbol{\theta}_M)$.
Cycle:
    $q_1^*(\boldsymbol{\theta}_1) \leftarrow \dfrac{\exp\{E_{-\boldsymbol{\theta}_1}\log p(\mathbf{y},\boldsymbol{\theta})\}}{\int\exp\{E_{-\boldsymbol{\theta}_1}\log p(\mathbf{y},\boldsymbol{\theta})\}\,d\boldsymbol{\theta}_1}$,
    $\vdots$
    $q_M^*(\boldsymbol{\theta}_M) \leftarrow \dfrac{\exp\{E_{-\boldsymbol{\theta}_M}\log p(\mathbf{y},\boldsymbol{\theta})\}}{\int\exp\{E_{-\boldsymbol{\theta}_M}\log p(\mathbf{y},\boldsymbol{\theta})\}\,d\boldsymbol{\theta}_M}$
until the increase in $\underline{p}(\mathbf{y};q)$ is negligible.
```

Then

$$\log\underline{p}(\mathbf{y};q) = \int q_1(\boldsymbol{\theta}_1)\log\left\{\frac{\widetilde{p}(\mathbf{y},\boldsymbol{\theta}_1)}{q(\boldsymbol{\theta}_1)}\right\}d\boldsymbol{\theta}_1$$
$$+ \text{ terms not involving } q_1.$$

By Result 1, the optimal $q_1$ is then

$$q_1^*(\boldsymbol{\theta}_1) = \widetilde{p}(\boldsymbol{\theta}_1|\mathbf{y}) \equiv \frac{\widetilde{p}(\mathbf{y},\boldsymbol{\theta}_1)}{\int\widetilde{p}(\mathbf{y},\boldsymbol{\theta}_1)\,d\boldsymbol{\theta}_1}$$
$$\propto \exp\left\{\int \log p(\mathbf{y},\boldsymbol{\theta})q_2(\boldsymbol{\theta}_2)\cdots q_M(\boldsymbol{\theta}_M)\,d\boldsymbol{\theta}_2\cdots d\boldsymbol{\theta}_M\right\}.$$

Repeating the same argument for maximizing $\log\underline{p}(\mathbf{y};q)$ over each of $q_2,\dots,q_M$ leads to the optimal densities satisfying:

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\{E_{-\boldsymbol{\theta}_i}\log p(\mathbf{y},\boldsymbol{\theta})\}, \qquad 1\le i\le M, \tag{5}$$

where $E_{-\boldsymbol{\theta}_i}$ denotes expectation with respect to the density $\prod_{j\neq i}q_j(\boldsymbol{\theta}_j)$. The iterative scheme, labeled Algorithm 1, can be used to solve for the $q_i^*$.

Convexity properties can be used to show that convergence to at least local optima is guaranteed (Boyd and Vandenberghe 2004). If conjugate priors are used, then the $q_i^*$ belong to recognizable density families and the $q_i^*$ updates reduce to updating parameters in the $q_i^*$ family (e.g., Winn and Bishop 2005). Also, in practice it is common to monitor convergence using $\log\{\underline{p}(\mathbf{y};q)\}$ rather than $\underline{p}(\mathbf{y};q)$. Sections 2.2.2–2.2.4 provide illustrations.

### *2.2.1 Connection With Gibbs Sampling*

It is easily shown that a valid alternative expression for the $q_i^*(\boldsymbol{\theta}_i)$ is

$$q_i^*(\boldsymbol{\theta}_i) \propto \exp\{E_{-\boldsymbol{\theta}_i}\log p(\boldsymbol{\theta}_i|\text{rest})\}, \tag{6}$$

where

$$\text{rest} \equiv \{\mathbf{y},\boldsymbol{\theta}_1,\dots,\boldsymbol{\theta}_{i-1},\boldsymbol{\theta}_{i+1},\dots,\boldsymbol{\theta}_M\}$$

is the set containing the random vectors in the model, apart from $\boldsymbol{\theta}_i$. The distributions $\boldsymbol{\theta}_i|\text{rest}$, $1\le i\le M$, are known, in the MCMC literature, as the *full conditionals*. This form of the optimal densities reveals a link with Gibbs sampling (e.g., Casella and George 1992) which involves successive draws from these full conditionals. Indeed, it becomes apparent from the upcoming examples that the product density transform approach leads to tractable solutions in situations where Gibbs sampling is also viable.

The DAG viewpoint of Bayesian models also gives rise to a useful result arising from the notion of *Markov blankets*. The Markov blanket of a node is the set of children, parents, and co-parents of that node. The result

$$p(\boldsymbol{\theta}_i|\text{rest}) = p(\boldsymbol{\theta}_i|\text{Markov blanket of }\boldsymbol{\theta}_i) \tag{7}$$

(Pearl 1988) means that determination of the required full conditionals involves localized calculations on the DAG. It follows from this fact and expression (6) that the product density approach involves a series of local operations. In Computer Science, this has become known as *variational message passing* (Winn and Bishop 2005). See the example in Section 2.2.3 for illustration of (7) and localization of variational updates.

### *2.2.2 Normal Random Sample*

Our first and most detailed illustration of variational approximation involves approximate Bayesian inference for the most familiar of statistical settings: a random sample from a Normal distribution. Specifically, consider

$$X_i|\mu,\sigma^2 \overset{\text{ind.}}{\sim} N(\mu,\sigma^2)$$

with priors

$$\mu\sim N(\mu_\mu,\sigma_\mu^2) \qquad\text{and}\qquad \sigma^2\sim \text{IG}(A,B).$$

The product density transform approximation to $p(\mu,\sigma^2|\mathbf{x})$ is

$$q(\mu,\sigma^2) = q_\mu(\mu)q_{\sigma^2}(\sigma^2). \tag{8}$$

The optimal densities take the form

$$q_\mu^*(\mu)\propto\exp[E_{\sigma^2}\{\log p(\mu|\sigma^2,\mathbf{x})\}] \qquad\text{and}$$
$$q_{\sigma^2}^*(\sigma^2)\propto\exp[E_\mu\{\log p(\sigma^2|\mu,\mathbf{x})\}],$$

where $\mathbf{x}=(X_1,\dots,X_n)$. Standard manipulations lead to the full conditionals being

$$\mu|\sigma^2,\mathbf{x}\sim N\left(\frac{n\overline{X}/\sigma^2+\mu_\mu/\sigma_\mu^2}{n/\sigma^2+1/\sigma_\mu^2},\frac{1}{n/\sigma^2+1/\sigma_\mu^2}\right) \qquad\text{and}$$
$$\sigma^2|\mu,\mathbf{x}\sim\text{IG}\left(A+\frac{n}{2},B+\frac{1}{2}\|\mathbf{x}-\mu\mathbf{1}_n\|^2\right),$$

where $\overline{X}=(X_1+\cdots+X_n)/n$ is the sample mean. The second of these, combined with (6), leads to

$$q_{\sigma^2}^*(\sigma^2)\propto\exp E_\mu\left\{-\left(A+\frac{n}{2}+1\right)\log(\sigma^2)\right.$$
$$\left.-\left(B+\frac{1}{2}\|\mathbf{x}-\mu\mathbf{1}_n\|^2\right)\Big/\sigma^2\right\}$$
$$\propto(\sigma^2)^{-(A+n/2+1)}$$
$$\times\exp\left\{-\left(B+\frac{1}{2}E_\mu\|\mathbf{x}-\mu\mathbf{1}_n\|^2\right)\Big/\sigma^2\right\}.$$

We recognize this as a member of the Inverse Gamma family:

$$q_{\sigma^2}^*(\sigma^2) \quad \text{is IG}\left(A+\frac{n}{2}, B+\frac{1}{2}E_\mu\|\mathbf{x}-\mu\mathbf{1}_n\|^2\right).$$

Note that $E_\mu\|\mathbf{x}-\mu\mathbf{1}_n\|^2 = \|\mathbf{x}-E_\mu(\mu)\mathbf{1}_n\|^2 + n\,\text{Var}_\mu(\mu)$ where

$$E_\mu(\mu) = \int_{-\infty}^{\infty} \mu_0 q_\mu(\mu_0)\,d\mu_0 \qquad \text{and}$$

$$\text{Var}_\mu(\mu) = \int_{-\infty}^{\infty} \{\mu_0 - E_\mu(\mu)\}^2 q_\mu(\mu_0)\,d\mu_0$$

are the mean and variance of the $q_\mu$ density. Similar arguments lead to

$$q_\mu^*(\mu) \quad \text{is } N\left(\frac{n\overline{X}E_{\sigma^2}(1/\sigma^2)+\mu_\mu/\sigma_\mu^2}{nE_{\sigma^2}(1/\sigma^2)+1/\sigma_\mu^2}, \frac{1}{nE_{\sigma^2}(1/\sigma^2)+1/\sigma_\mu^2}\right), \tag{9}$$

where $E_{\sigma^2}(1/\sigma^2) = \int_0^\infty (1/\sigma_0^2) q_{\sigma^2}(\sigma_0^2)\,d\sigma_0^2$. When $q_{\sigma^2} = q_{\sigma^2}^*$ we get

$$E_{\sigma^2}(1/\sigma^2) = \frac{A+n/2}{B+\frac{1}{2}\{\|\mathbf{x}-E_\mu(\mu)\mathbf{1}_n\|^2 + n\,\text{Var}_\mu(\mu)\}}. \tag{10}$$

It is now apparent that the functional forms of the optimal densities $q_\mu^*$ and $q_{\sigma^2}^*$ are Normal and Inverse Gaussian, respectively, but the parameters need to be determined from relationships such as (9) and (10). Let

$$\mu_{q(\mu)} \equiv E_\mu(\mu), \qquad \sigma_{q(\mu)}^2 \equiv \text{Var}_\mu(\mu), \qquad \text{and}$$

$$B_{q(\sigma^2)} \equiv \left(A+\frac{n}{2}\right)\Big/E_{\sigma^2}(1/\sigma^2).$$

Using the relationships established at (9) and (10) we arrive at Algorithm 2, which can be used to obtain the optimal values of $\mu_{q(\mu)}$, $\sigma_{q(\mu)}^2$, and $B_{q(\sigma^2)}$.

Note that $\log \underline{p}(\mathbf{x}; q)$ admits the explicit expression:

$$\log \underline{p}(\mathbf{x}; q) = \frac{1}{2} - \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\left(\sigma_{q(\mu)}^2/\sigma_\mu^2\right) - \frac{(\mu_{q(\mu)}-\mu_\mu)^2 + \sigma_{q(\mu)}^2}{2\sigma_\mu^2}$$

$$+ A\log(B) - \left(A+\frac{n}{2}\right)\log\left(B_{q(\sigma^2)}\right) + \log\Gamma\left(A+\frac{n}{2}\right) - \log\Gamma(A).$$

**Algorithm 2** Iterative scheme for obtaining the parameters in the optimal densities $q_\mu^*$ and $q_{\sigma^2}^*$ in the Normal random sample example.

```
Initialize: $B_{q(\sigma^2)} > 0$.
Cycle:
    $\sigma_{q(\mu)}^2 \leftarrow \left\{n\left(A+\frac{n}{2}\right)\Big/B_{q(\sigma^2)} + 1/\sigma_\mu^2\right\}^{-1}$,
    $\mu_{q(\mu)} \leftarrow \left\{n\overline{X}\left(A+\frac{n}{2}\right)\Big/B_{q(\sigma^2)} + \mu_\mu/\sigma_\mu^2\right\}\sigma_{q(\mu)}^2$,
    $B_{q(\sigma^2)} \leftarrow B + \frac{1}{2}\left(\|\mathbf{x}-\mu_{q(\mu)}\mathbf{1}_n\|^2 + n\sigma_{q(\mu)}^2\right)$
until the increase in $\underline{p}(\mathbf{x}; q)$ is negligible.
```

However, within each iteration of Algorithm 2, this expression is valid only after each of the parameter updates has been made.

Upon convergence to $\mu_{q(\mu)}^*$, $(\sigma_{q(\mu)}^2)^*$, and $B_{q(\sigma^2)}^*$, the approximations to the individual posterior densities are

$$p(\mu|\mathbf{x}) \approx \left\{2\pi(\sigma_{q(\mu)}^2)^*\right\}^{-1/2} \exp\left[-\left(\mu-\mu_{q(\mu)}^*\right)^2/\left\{2(\sigma_{q(\mu)}^2)^*\right\}\right]$$

and

$$p(\sigma^2|\mathbf{x}) \approx \frac{(B_{q(\sigma^2)}^*)^{A+n/2}}{\Gamma(A+\frac{n}{2})}(\sigma^2)^{-A-n/2-1} \times \exp\left(-B_{q(\sigma^2)}^*/\sigma^2\right), \qquad \sigma^2 > 0.$$

Figure 2 illustrates these variational approximations for a simulated sample of size $n = 20$ from the $N(100, 225)$ density. For priors we used $\mu \sim N(0, 10^8)$ and $\sigma^2 \sim \text{IG}(\frac{1}{100}, \frac{1}{100})$, corresponding to vague beliefs about the mean and variance, and such that the prior mean of the precision, $1/\sigma^2$, is unity. The initial value for the iterative scheme is $B_{q(\sigma^2)} = 1$. The exact posterior densities, obtained via highly accurate quadrature, are also displayed. Note that, in this example, convergence is very rapid and the accuracy of the variational approximation is quite good.

### 2.2.3 *Linear Mixed Model*

The Bayesian version of the Gaussian linear mixed model takes the general form

$$\mathbf{y}|\boldsymbol{\beta}, \mathbf{u}, \mathbf{G}, \mathbf{R} \sim N(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}, \mathbf{R}), \qquad \mathbf{u}|\mathbf{G} \sim N(\mathbf{0}, \mathbf{G}), \tag{11}$$

where $\mathbf{y}$ is an $n \times 1$ vector of response variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are corresponding design matrices, and $\mathbf{G}$ and $\mathbf{R}$ are covariance matrices. While several possibilities exist for $\mathbf{G}$ and $\mathbf{R}$ (e.g., McCulloch, Searle, and Neuhaus 2008), we restrict attention here to *variance component* models with

$$\mathbf{G} = \text{blockdiag}\left(\sigma_{u1}^2\mathbf{I}_{K_1}, \ldots, \sigma_{ur}^2\mathbf{I}_{K_r}\right) \quad \text{and} \quad \mathbf{R} = \sigma_\varepsilon^2\mathbf{I}. \tag{12}$$

We also impose the conjugate priors:

$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2\mathbf{I}), \qquad \sigma_{u\ell}^2 \sim \text{IG}(A_{u\ell}, B_{u\ell}), \quad 1 \le \ell \le r, \qquad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon) \tag{13}$$

for some $\sigma_\beta^2, A_{u\ell}, B_{u\ell}, A_\varepsilon, B_\varepsilon > 0$. Figure 3 is the DAG corresponding to model (11)–(13).

Somewhat remarkably, a tractable solution arises for the two-component product

$$q(\boldsymbol{\beta}, \mathbf{u}, \sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2) = q_{\boldsymbol{\beta},\mathbf{u}}(\boldsymbol{\beta}, \mathbf{u})q_{\sigma^2}(\sigma_{u1}^2, \ldots, \sigma_{ur}^2, \sigma_\varepsilon^2). \tag{14}$$

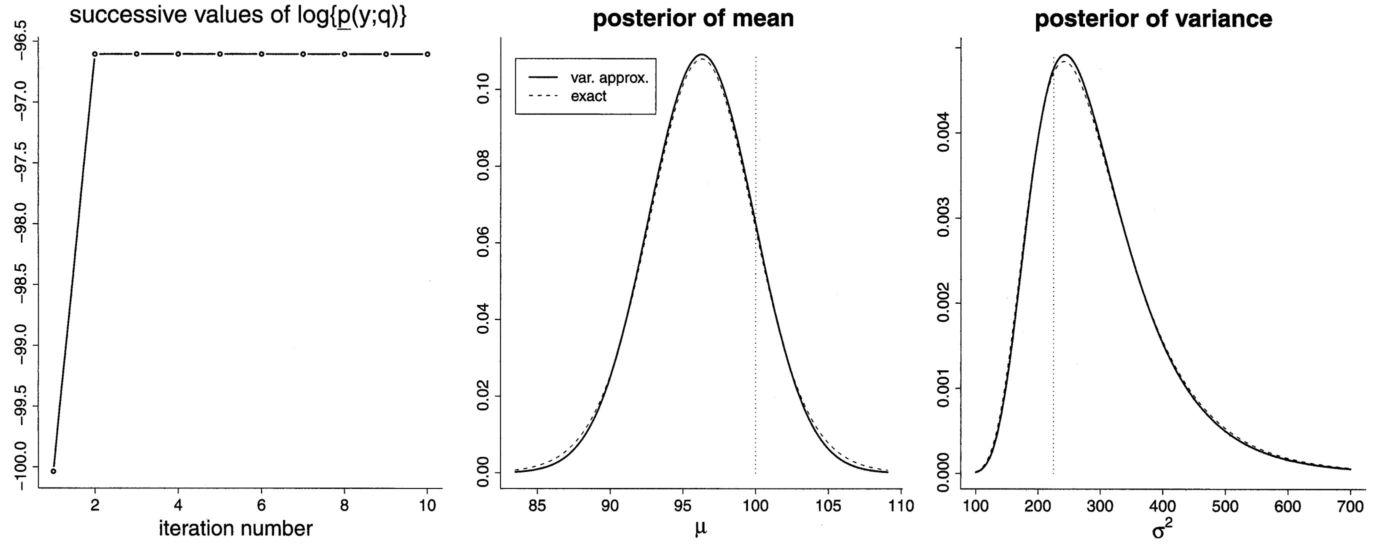

Figure 2. Results from applying the product density variational approximation to a simulated Normal random sample. The exact posterior density functions are added for comparison. The vertical dotted line in the posterior density plots corresponds to the true value of the parameter.

Application of (5) leads to the optimal densities taking the form

$q^*_{\boldsymbol{\beta},\mathbf{u}}(\boldsymbol{\beta},\mathbf{u})$ is a Multivariate Normal density function,

$q^*_{\sigma^2}$ is a product of $r+1$ Inverse Gamma density functions.

It should be stressed that these forms are not imposed at the outset, but arise as optimal solutions for model (11)–(13) and product restriction (14). Moreover, the factorization of $q^*_{\sigma^2}$ into $r+1$ separate components is also a consequence of (5) for the current model, rather than an imposition. Bishop (2006, sec. 10.2.5) explained how these *induced* factorizations follow from the structure of the DAG and d-separation theory (Pearl 1988). This example also benefits from the Markov blanket result (7) described in Section 2.2.1 and Figure 3. For example, the full conditional density of $\sigma^2_{u1}$ is

$$\begin{aligned}p(\sigma^2_{u1}|\text{rest}) &= p(\sigma^2_{u1}|\text{Markov blanket of } \sigma^2_{u1})\\ &= p(\sigma^2_{u1}|\mathbf{u},\sigma^2_{u2},\ldots,\sigma^2_{ur}).\end{aligned}$$

Hence, determination of $q^*_{\sigma^2_{u1}}$ requires calculations involving only the subset of the DAG consisting of $\mathbf{u}$ and the variance parameters.

Let $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})}$ be the mean and covariance matrix for the $q^*_{\boldsymbol{\beta},\mathbf{u}}$ density and set $\mathbf{C}\equiv[\mathbf{X}\ \mathbf{Z}]$. For the $q^*_{\sigma^2}$ density the shape parameters for the $r+1$ components can be shown to be deterministic: $A_{u1}+\frac{1}{2}K_1,\ldots,A_{ur}+\frac{1}{2}K_r$, $A_\varepsilon+\frac{1}{2}n$. Let $B_{q(\sigma^2_{u1})},\ldots,B_{q(\sigma^2_{ur})},B_{q(\sigma^2_\varepsilon)}$ be the accompanying rate parameters. The relationships between $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})})$ and $(B_{q(\sigma^2_{u1})},\ldots,B_{q(\sigma^2_{ur})},B_{q(\sigma^2_\varepsilon)})$ enforced by (5) lead to the iterative scheme in Algorithm 3.

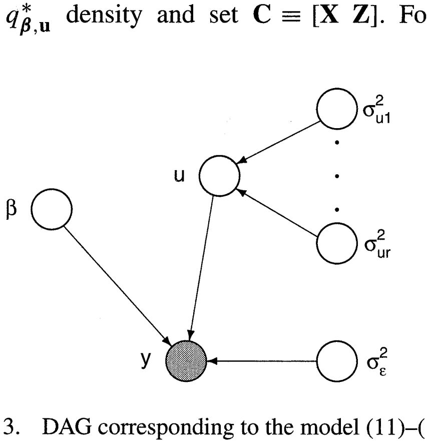

Figure 3. DAG corresponding to the model (11)–(13).

In this case $\log\underline{p}(\mathbf{y};q)$ takes the form

$$\begin{aligned}&\log\underline{p}(\mathbf{y};q)\\ &\quad=\frac{1}{2}\left(p+\sum_{\ell=1}^{r}K_\ell\right)-\frac{n}{2}\log(2\pi)-\frac{p}{2}\log(\sigma^2_{\beta})\\ &\qquad+\frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})}|-\frac{1}{2\sigma^2_\beta}\left\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\operatorname{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\right\}\\ &\qquad+A_\varepsilon\log(B_\varepsilon)-\left(A_\varepsilon+\frac{n}{2}\right)\log(B_{q(\sigma^2_\varepsilon)})\\ &\qquad+\log\Gamma\left(A_\varepsilon+\frac{n}{2}\right)-\log\Gamma(A_\varepsilon)\\ &\qquad+\sum_{\ell=1}^{r}\left\{A_{u\ell}\log(B_{u\ell})-\left(A_{u\ell}+\frac{K_\ell}{2}\right)\log(B_{q(\sigma^2_{u\ell})})\right.\\ &\qquad\left.+\log\Gamma\left(A_{u\ell}+\frac{K_\ell}{2}\right)-\log\Gamma(A_{u\ell})\right\}.\end{aligned}$$

Note that, within each iteration of Algorithm 3, this expression applies only after each of the parameter updates has been made.

Upon convergence to $\boldsymbol{\mu}^*_{q(\boldsymbol{\beta},\mathbf{u})},\boldsymbol{\Sigma}^*_{q(\boldsymbol{\beta},\mathbf{u})},B^*_{q(\sigma^2_{u1})},\ldots,B^*_{q(\sigma^2_{ur})}$ and $B^*_{q(\sigma^2_\varepsilon)}$ the approximate posteriors are:

$$p(\boldsymbol{\beta},\mathbf{u}|\mathbf{y})\approx\text{the } N(\boldsymbol{\mu}^*_{q(\boldsymbol{\beta},\mathbf{u})},\boldsymbol{\Sigma}^*_{q(\boldsymbol{\beta},\mathbf{u})})\text{ density function}$$

and

$$\begin{aligned}&p(\sigma^2_{u1},\ldots,\sigma^2_{ur},\sigma^2_\varepsilon|\mathbf{y})\\ &\quad\approx\text{product of the IG}(A_{u\ell}+\tfrac{1}{2}K_\ell,B^*_{q(\sigma^2_{u\ell})}),\ 1\le\ell\le r,\end{aligned}$$

**Algorithm 3** Iterative scheme for obtaining the parameters in the optimal densities $q^*_{\boldsymbol{\beta},\mathbf{u}}$ and $q^*_{\sigma^2}$ in the Bayesian linear mixed model example.

```
Initialize: $B_{q(\sigma_\varepsilon^2)}, B_{q(\sigma_{u1}^2)}, \ldots, B_{q(\sigma_{ur}^2)} > 0$.
Cycle:
    $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})} \leftarrow \left\{ \frac{A_\varepsilon + \frac{n}{2}}{B_{q(\sigma_\varepsilon^2)}} \mathbf{C}^T\mathbf{C} + \mathrm{blockdiag}\left( \sigma_\beta^{-2}\mathbf{I}_p, \frac{A_{u1} + \frac{1}{2}K_1}{B_{q(\sigma_{u1}^2)}} \mathbf{I}_{K_1}, \ldots, \frac{A_{ur} + \frac{1}{2}K_r}{B_{q(\sigma_{ur}^2)}} \mathbf{I}_{K_r} \right) \right\}^{-1}$,
    $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u})} \leftarrow \left( \frac{A_\varepsilon + \frac{n}{2}}{B_{q(\sigma_\varepsilon^2)}} \right) \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})} \mathbf{C}^T \mathbf{y}$,
    $B_{q(\sigma_\varepsilon^2)} \leftarrow B_\varepsilon + \frac{1}{2}\left\{ \left\| \mathbf{y} - \mathbf{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\mathbf{u})} \right\|^2 + \mathrm{tr}\left( \mathbf{C}^T\mathbf{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\mathbf{u})} \right) \right\}$,
    $B_{q(\sigma_{u\ell}^2)} \leftarrow B_{u\ell} + \frac{1}{2}\left\{ \left\| \boldsymbol{\mu}_{q(\mathbf{u}_\ell)} \right\|^2 + \mathrm{tr}\left( \boldsymbol{\Sigma}_{q(\mathbf{u}_\ell)} \right) \right\}$  for $1 \le \ell \le r$
until the increase in $\underline{p}(\mathbf{y}; q)$ is negligible.
```

density functions together with the $\mathrm{IG}(A_\varepsilon + \frac{1}{2}n, B^*_{q(\sigma_\varepsilon^2)})$ density function.

We now provide an illustration for Bayesian analysis of a dataset involving longitudinal orthodontic measurements on 27 children (source: Pinheiro and Bates 2000). The data are available in the R computing environment (R Development Core Team 2010) via the package nlme (Pinheiro et al. 2009), in the object Orthodont. We entertained the random intercept model

$$\begin{aligned} \mathtt{distance}_{ij} | U_i &\overset{\text{ind.}}{\sim} N(\beta_0 + U_i + \beta_1 \mathtt{age}_{ij} \\ &\qquad + \beta_2 \mathtt{male}_i, \sigma_\varepsilon^2), \\ U_i | \sigma_u^2 &\overset{\text{ind.}}{\sim} N(0, \sigma_u^2), \qquad 1 \le i \le 27, 1 \le j \le 4, \\ \beta_i &\overset{\text{ind.}}{\sim} N(0, \sigma_\beta^2), \qquad \sigma_u^2, \sigma_\varepsilon^2 \overset{\text{ind.}}{\sim} \mathrm{IG}(A, B), \end{aligned} \tag{15}$$

where $\mathtt{distance}_{ij}$ is the distance from the pituitary to the pterygomaxillary fissure (mm) for patient $i$ at time point $j$. Similarly, $\mathtt{age}_{ij}$ correspond to the longitudinal age values in years and $\mathtt{male}_i$ is an indicator of the $i$th child being male. This fits into framework (11)–(12) with $\mathbf{y}$ containing the $\mathtt{distance}_{ij}$ measurement, $\mathbf{X} = [1, \mathtt{age}_{ij}, \mathtt{male}_i]$, and $\mathbf{Z} = \mathbf{I}_{27} \otimes \mathbf{1}_4$ is an indicator matrix for the random intercepts. We used the vague priors $\sigma_\beta^2 = 10^8$, $A = B = \frac{1}{100}$ and used standardized versions of the distance and age data during the fitting. The results were then converted back to the original units. For comparison, we obtained 1 million samples from the posteriors using MCMC (with a burn-in of length 5000) and, from these, constructed kernel density estimate approximations to the posteriors. For such a high Monte Carlo sample size we would expect these MCMC-based approximations to be very accurate.

Figure 4 shows the progressive values of $\log \underline{p}(\mathbf{y}; q)$ and the approximate posterior densities obtained from applying Algorithm 3. Once again, convergence of $\log\{\underline{p}(\mathbf{y}; q)\}$ to a maximum is seen to be quite rapid. The variational approximate posterior densities are quite close to those obtained via MCMC, and indicate statistical significance of all model parameters in the sense that most of the posterior probability mass is away from zero.

### 2.2.4 *Probit Regression and the Use of Auxiliary Variables*

As shown by Albert and Chib (1993), Gibbs sampling for the Bayesian probit regression model becomes tractable when a particular set of auxiliary variables is introduced. The same trick applies to product density variational approximation (Girolami and Rogers 2006; Consonni and Marin 2007), as we now show.

The Bayesian probit regression model that we consider here is

$$\begin{aligned} &Y_i | \beta_0, \ldots, \beta_k \\ &\overset{\text{ind.}}{\sim} \mathrm{Bernoulli}(\Phi(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})), \qquad 1 \le i \le n, \end{aligned}$$

where the prior distribution on the coefficient vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ takes the form $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. Letting $\mathbf{X} \equiv [1 \ x_{1i} \ \cdots \ x_{ki}]_{1 \le i \le n}$, the likelihood can be written compactly as

$$p(\mathbf{y}|\boldsymbol{\beta}) = \Phi(\mathbf{X}\boldsymbol{\beta})^{\mathbf{y}} \{\mathbf{1}_n - \Phi(\mathbf{X}\boldsymbol{\beta})\}^{\mathbf{1}_n - \mathbf{y}}, \qquad \boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}).$$

Introduce the vector of auxiliary variables $\mathbf{a} = (a_1, \ldots, a_n)$, where

$$a_i | \boldsymbol{\beta} \overset{\text{ind.}}{\sim} N((\mathbf{X}\boldsymbol{\beta})_i, 1).$$

This allows us to write

$$p(y_i|a_i) = I(a_i \ge 0)^{y_i} I(a_i < 0)^{1-y_i}, \qquad 1 \le i \le n.$$

In graphical model terms we are introducing a new node to the graph, as conveyed by Figure 5. Expansion of the parameter set from $\{\boldsymbol{\beta}\}$ to $\{\mathbf{a}, \boldsymbol{\beta}\}$ is the key to achieving a tractable solution.

Consider the product restriction

$$q(\mathbf{a}, \boldsymbol{\beta}) = q_{\mathbf{a}}(\mathbf{a}) q_{\boldsymbol{\beta}}(\boldsymbol{\beta}).$$

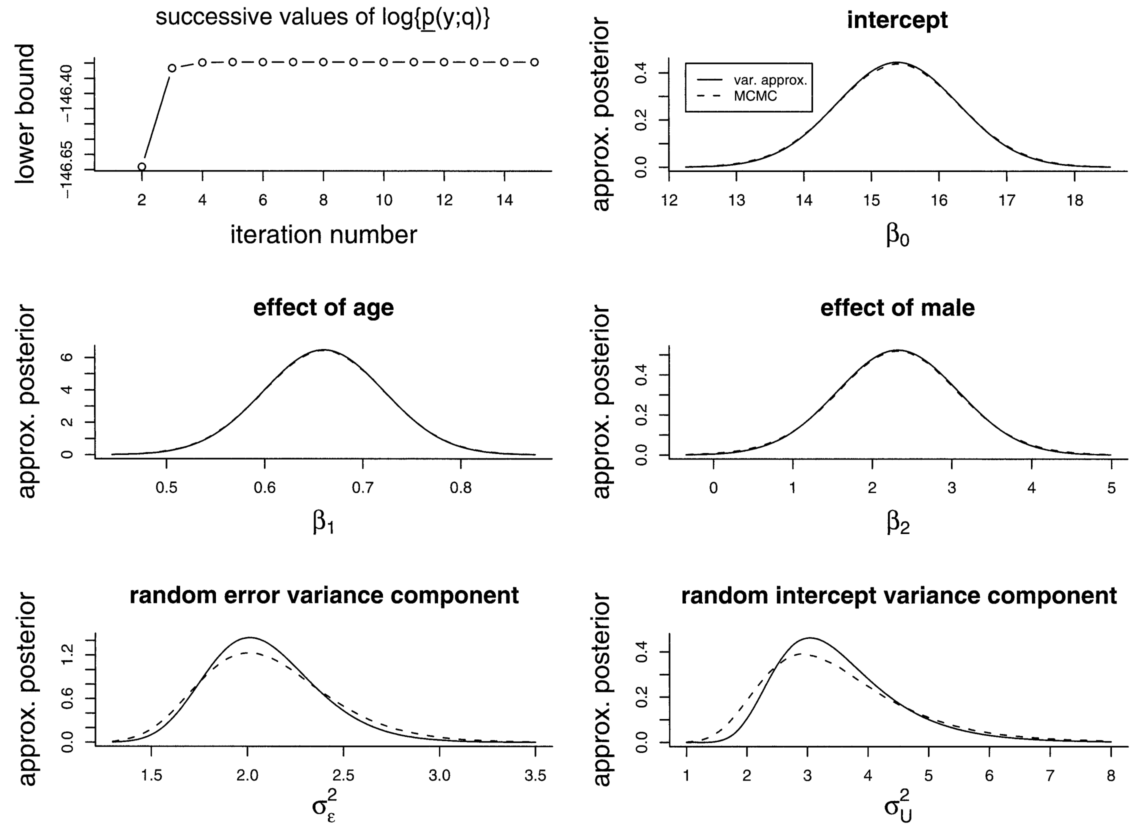

Figure 4. Approximate posterior densities from applying the product density variational approximation to (11)–(13) for the orthodontic data. 'Exact' posterior densities, based on kernel density estimates of 1 million MCMC samples, are shown for comparison.

Then application of (5) leads to

$$q^*_{\mathbf{a}}(\mathbf{a}) = \left[\prod_{i=1}^{n}\left\{\frac{I(a_i \ge 0)}{\Phi((\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})_i)}\right\}^{y_i}\left\{\frac{I(a_i < 0)}{1-\Phi((\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})_i)}\right\}^{1-y_i}\right]$$
$$\times (2\pi)^{-n/2}\exp\left\{-\frac{1}{2}\|\mathbf{a}-\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2\right\}$$

and $q^*_{\boldsymbol{\beta}}(\boldsymbol{\beta})$ is the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, (\mathbf{X}^T\mathbf{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1})$ density function. These optimal densities are specified up to the parameter vector $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \equiv E_{\boldsymbol{\beta}}(\boldsymbol{\beta})$. We also need to work with the $q$-density mean of the auxiliary variable vector $\boldsymbol{\mu}_{q(\mathbf{a})} \equiv E_{\mathbf{a}}(\mathbf{a})$. The iterative scheme, Algorithm 4, emerges.

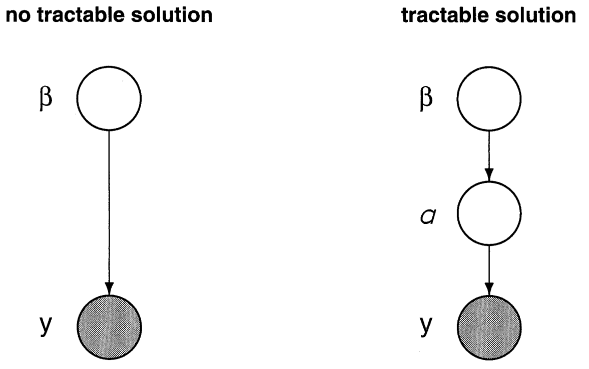

Figure 5. Graphical representations of the probit regression model. The left graph does not admit a tractable product density variational approximation. The right graph overcomes this with the addition of an auxiliary variable node.

The $\log \underline{p}(\mathbf{y}; q)$ expression in this case is

$$\begin{aligned}\log \underline{p}(\mathbf{y}; q) &= \mathbf{y}^T \log\{\Phi(\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})\}\\ &\quad + (\mathbf{1}_n - \mathbf{y})^T \log\{\mathbf{1}_n - \Phi(\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})\}\\ &\quad - \frac{1}{2}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\\ &\quad - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\mathbf{X}^T\mathbf{X} + \mathbf{I}|.\end{aligned}$$

**Algorithm 4** Iterative scheme for obtaining the parameters in the optimal densities $q^*_{\boldsymbol{\beta}}$ and $q^*_{\mathbf{a}}$ in the Bayesian probit regression example.

```
Initialize: $\boldsymbol{\mu}_{q(\mathbf{a})} (n \times 1)$.
Cycle:
    $\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \leftarrow (\mathbf{X}^T\mathbf{X}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}(\mathbf{X}^T\boldsymbol{\mu}_{q(\mathbf{a})}+\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}})$,
    $\boldsymbol{\mu}_{q(\mathbf{a})} \leftarrow \mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \frac{\phi(\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})}{\Phi(\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})^{\mathbf{y}}\{\Phi(\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})})-\mathbf{1}_n\}^{\mathbf{1}_n-\mathbf{y}}}$
until the increase in $\underline{p}(\mathbf{y}; q)$ is negligible.
```

Upon convergence, the approximate posterior distribution of the regression coefficients is

$$\boldsymbol{\beta}|\mathbf{y} \overset{\text{approx.}}{\sim} N\big(\boldsymbol{\mu}^*_{q(\boldsymbol{\beta})}, (\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1})^{-1}\big).$$

### 2.2.5 *Finite Normal Mixture Model*

Our last example of product density variational approximation is of great interest within both Statistics and Computer Science: inference for finite mixture models. Let $X_1, \ldots, X_n$ be a univariate sample that is modeled as a random sample from a mixture of $K$ Normal density functions with parameters $(\mu_k, \sigma_k^2)$, $1 \le k \le K$. Accordingly, the joint density function of the sample is

$$\begin{aligned} p(x_1, \ldots, x_n) \\ = \prod_{i=1}^{n} \left[ \sum_{k=1}^{K} w_k (2\pi\sigma_k^2)^{-1/2} \exp\left\{ -\frac{1}{2}(x_i - \mu_k)^2/\sigma_k^2 \right\} \right], \quad (16) \end{aligned}$$

where the weights $w_k$, $1 \le k \le K$, are nonnegative, and sum to unity. Let $(w_1, \ldots, w_K)$ have prior distribution:

$$(w_1, \ldots, w_K) \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \qquad \alpha > 0.$$

We will take the prior distributions for the mean and variance parameters to be

$$\mu_k \overset{\text{ind.}}{\sim} N\big(\mu_{\mu_k}, \sigma^2_{\mu_k}\big), \qquad \sigma_k^2 \overset{\text{ind.}}{\sim} \text{IG}(A_k, B_k), \qquad 1 \le k \le K.$$

As with the probit regression model, a tractable product density transform requires the introduction of the auxiliary variable vectors:

$$\begin{aligned} &(a_{i1}, \ldots, a_{ik})|(w_1, \ldots, w_K) \\ &\overset{\text{ind.}}{\sim} \text{Multinomial}(1; w_1, \ldots, w_K), \qquad 1 \le i \le n. \quad (17) \end{aligned}$$

According to this notation, $\sum_{k=1}^K a_{ik} = 1$ and $w_k = P(a_{ik} = 1)$. If we set

$$\begin{aligned} &p(x_i | a_{i1}, \ldots, a_{iK}) \\ &= \prod_{k=1}^{K} \left[ (2\pi\sigma_k^2)^{-1/2} \exp\left\{ -\frac{1}{2}(x_i - \mu_k)^2/\sigma_k^2 \right\} \right]^{a_{ik}} \end{aligned}$$

independently for each $1 \le i \le n$, then, using (17), the joint density function of the $X_1, \ldots, X_n$ is easily shown to be (16).

Let $\mathbf{w}$, $\boldsymbol{\mu}$, $\boldsymbol{\sigma}^2$, and $\mathbf{a}$ be the vectors containing the corresponding subscripted random variables. Then either of the product density restrictions

$$\begin{aligned} q(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{a}) &= q(\mathbf{w}, \boldsymbol{\mu}) q(\boldsymbol{\sigma}^2) q(\mathbf{a}) \qquad \text{or} \\ q(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{a}) &= q(\mathbf{w}, \boldsymbol{\sigma}^2) q(\boldsymbol{\mu}) q(\mathbf{a}) \end{aligned} \qquad (18)$$

is sufficient for a closed form solution. Note that subscripting on the $q$ densities is being suppressed to reduce clutter. Regardless of which restriction in (18) is chosen, application of Algorithm 1 leads to the optimal density for the model parameters having the product structure

$$q^*(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) = q^*(\mathbf{w}) q^*(\boldsymbol{\mu}) q^*(\boldsymbol{\sigma}^2),$$

where

$$\begin{aligned} q^*(\mathbf{w}) &= \text{density function of a Dirichlet distribution,} \\ q^*(\boldsymbol{\mu}) &= \text{product of } K \text{ Normal density functions,} \end{aligned}$$

and

$$q^*(\boldsymbol{\sigma}^2) = \text{product of } K \text{ Inverse Gamma density functions.}$$

For $1 \le k \le K$, let $\mu_{q(\mu_k)}$ and $\sigma^2_{q(\mu_k)}$ denote the mean and variance for $q^*(\mu_k)$ and let $A_{q(\sigma_k^2)}$ and $B_{q(\sigma_k^2)}$ denote the shape and rate parameters for $q^*(\sigma_k^2)$. Also, let

$$\boldsymbol{\alpha}_{q(\mathbf{w})} \equiv \big(\alpha_{q(w_1)}, \ldots, \alpha_{q(w_1)}\big)$$

be the Dirichlet parameter vector for $q^*(\mathbf{w})$. The optimal parameters may be found obtained using Algorithm 5. Recall, from Section 1.1, that $\psi$ denotes the Digamma function.

The $\log \underline{p}(\mathbf{x}; q)$ expression in this case is

$$\begin{aligned} \log \underline{p}(\mathbf{x}; q) = {} & \frac{1}{2}K\{1 - n\log(2\pi)\} + \log\Gamma(K\alpha) \\ & - K\log\Gamma(\alpha) - \log\Gamma(n + K\alpha) \\ & + \sum_{k=1}^{K} \Bigg[ A_k \log(B_k) - A_{q(\sigma_k^2)} \log\big(B_{q(\sigma_k^2)}\big) \\ & + \log\Gamma\big(A_{q(\sigma_k^2)}\big) - \log\Gamma(A_k) \\ & + \log\Gamma\big(\alpha_{q(w_k)}\big) + \frac{1}{2}\log\big(\sigma^2_{q(\mu_k)}/\sigma^2_{\mu_k}\big) \\ & - \frac{1}{2}\big\{\big(\mu_{q(\mu_k)} - \mu_{\mu_k}\big)^2 + \sigma^2_{q(\mu_k)}\big\}/\sigma^2_{\mu_k} \\ & - \sum_{i=1}^{n} \omega_{ik} \log(\omega_{ik}) \Bigg]. \end{aligned}$$

Note that, for each iteration of Algorithm 5, this expression is valid only after each of the parameter updates has been made.

Algorithm 5 is similar to the EM algorithm for fitting a finite Normal mixture model. Comparison and contrast are given in section 10.2.1 of the book by Bishop (2006).

Figure 6 shows the result of applying Algorithm 5 to data on the duration of geyser eruptions. The data are available in the R computing environment via the package MASS (Venables and Ripley 2009), in the object `geyser$duration`. The number of mixtures was set at $K = 2$ and vague priors with $\alpha = 0.001$, $\mu_k \sim N(0, 10^8)$, and $\sigma_k^2 \sim \text{IG}(\frac{1}{100}, \frac{1}{100})$ were used. The upper panel of Figure 6 shows that convergence of $\log \underline{p}(\mathbf{x}; q)$ was obtained after about 20 iterations from naïve starting values. In the lower panel, the curve corresponds to the approximate posterior mean of the common density function. The shaded region corresponds to approximate pointwise 95% credible sets. These were obtained using 10,000 draws from $q^*(\mathbf{w}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2)$.

Finally, we note that variational approximation methodology could also be used to choose the number of mixtures $K$. See, for example, the work of Bishop (2006, sec. 10.2.4) and McGrory and Titterington (2007).

**Algorithm 5** Iterative scheme for obtaining the parameters in the optimal densities $q^*_{\mathbf{w}}$, $q^*_{\boldsymbol{\mu}}$, and $q^*_{\boldsymbol{\sigma}^2}$ in the finite Normal mixtures example.

```
Initialize: $\mu_{q(\mu_k)} \in \mathbb{R}$ and $\boldsymbol{\alpha}_{q(w_k)}, \sigma^2_{q(\mu_k)}, A_{q(\sigma^2_k)}, B_{q(\sigma^2_k)}, \omega_{\bullet k} > 0$, $1 \le k \le K$,
        such that $\sum_{k=1}^K \omega_{\bullet k} = 1$.
Cycle: For $i = 1, \dots, n$ and $k = 1, \dots, K$:
        $\nu_{ik} \leftarrow \psi(\alpha_{q(w_k)}) + \frac{1}{2}\psi(A_{q(\sigma^2_k)}) - \frac{1}{2}\log(B_{q(\sigma^2_k)}) - \frac{1}{2}A_{q(\sigma^2_k)}\{(X_i - \mu_{q(\mu_k)})^2 + \sigma^2_{q(\mu_k)}\}/B_{q(\sigma^2_k)}$.
    For $i = 1, \dots, n$ and $k = 1, \dots, K$: $\omega_{ik} \leftarrow \exp(\nu_{ik})/\sum_{k=1}^K \exp(\nu_{ik})$.
    For $k = 1, \dots, K$:
        $\omega_{\bullet k} \leftarrow \sum_{i=1}^n \omega_{ik}$;    $\sigma^2_{q(\mu_k)} \leftarrow 1/\{1/\sigma^2_{\mu_k} + A_{q(\sigma^2_k)}\omega_{\bullet k}/B_{q(\sigma^2_k)}\}$,
        $\mu_{q(\mu_k)} \leftarrow \sigma^2_{q(\mu_k)}\left\{\mu_{\mu_k}/\sigma^2_{\mu_k} + A_{q(\sigma^2_k)}\sum_{i=1}^n \omega_{ik}X_i/B_{q(\sigma^2_k)}\right\}$,
        $\alpha_{q(w_k)} \leftarrow \alpha + \omega_{\bullet k}$;    $A_{q(\sigma^2_k)} \leftarrow A_k + \frac{1}{2}\omega_{\bullet k}$,
        $B_{q(\sigma^2_k)} \leftarrow B_k + \frac{1}{2}\sum_{i=1}^n \omega_{ik}\{(X_i - \mu_{q(\mu_k)})^2 + \sigma^2_{q(\mu_k)}\}$
until the increase in $\underline{p}(\mathbf{x}; q)$ is negligible.
```

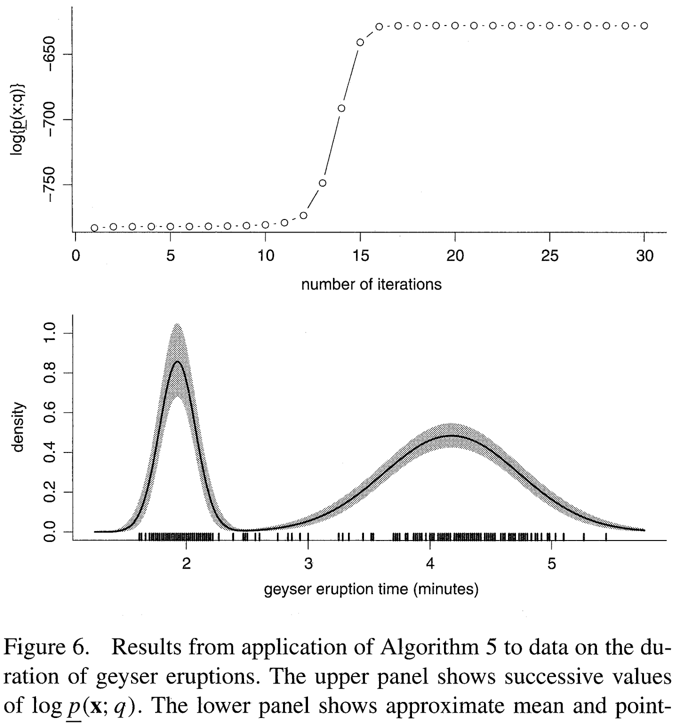

Figure 6. Results from application of Algorithm 5 to data on the duration of geyser eruptions. The upper panel shows successive values of $\log \underline{p}(\mathbf{x}; q)$. The lower panel shows approximate mean and pointwise 95% credible sets for the common density function. The data are shown at the base of the plot.

### 2.3 Parametric Density Transforms

Rather than assuming that $q(\boldsymbol{\theta})$ has product density structure, we may instead assume that it belongs to a particular parametric family and hope that this results in a more tractable approximation to the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. This approach has received less attention in the Computer Science literature. Examples where it has appeared are the works by Barber and Bishop (1998), Seeger (2000, 2004), Honkela and Valpola (2005), and Archambeau et al. (2007).

Next, we illustrate parametric density transforms with a simple example.

#### 2.3.1 *Poisson Regression With Gaussian Transform*

Consider the Bayesian Poisson regression model

$$Y_i|\beta_0, \dots, \beta_k \overset{\text{ind.}}{\sim} \text{Poisson}(\exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki})), \qquad 1 \le i \le n,$$

where the prior distribution on the coefficient vector $\boldsymbol{\beta} \equiv (\beta_0, \dots, \beta_k)$ takes the form $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. As before, we let $\mathbf{X} = [1 \ x_{1i} \ \cdots \ x_{ki}]_{1 \le i \le n}$. Then the likelihood is

$$p(\mathbf{y}|\boldsymbol{\beta}) = \exp\{\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T \exp(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}_n^T \log(\mathbf{y}!)\}$$

and the marginal likelihood is

$$p(\mathbf{y}) = (2\pi)^{-(k+1)/2}|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|^{-1/2} \times \int_{\mathbb{R}^{k+1}} \exp\left\{\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T \exp(\mathbf{X}\boldsymbol{\beta}) - \mathbf{1}_n^T \log(\mathbf{y}!) - \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\} d\boldsymbol{\beta}.$$

Note that $p(\mathbf{y})$, and hence $p(\boldsymbol{\beta}|\mathbf{y})$, involves an intractable integral over $\mathbb{R}^{k+1}$.

Take $q$ to be the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$ density:

$$\begin{aligned} q(\boldsymbol{\beta}; \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) &\\ = (2\pi)^{-p/2}\left|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\right|^{-1/2} &\\ \times \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})^T \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{q(\boldsymbol{\beta})})\right\}. \end{aligned}$$

Then the lower bound (4) admits the explicit expression

$$\begin{aligned} \log \underline{p}(\mathbf{y}; \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) &\\ = \mathbf{y}^T\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \mathbf{1}_n^T \exp\left\{\mathbf{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} + \frac{1}{2}\,\text{diagonal}(\mathbf{X}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\mathbf{X}^T)\right\} &\\ - \frac{1}{2}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{\mu}_{\boldsymbol{\beta}}) - \frac{1}{2}\text{tr}(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}) &\\ + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| + \frac{k+1}{2} &\\ - \mathbf{1}_n^T \log(\mathbf{y}!). &\qquad (19) \end{aligned}$$

Note that, from (2),

$$\log p(\mathbf{y}) \geq \log \underline{p}(\mathbf{y}; \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$$

for all choices of the mean vector $\boldsymbol{\mu}_{q(\boldsymbol{\beta})}$ and covariance matrix $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}$. Choosing these *variational parameters* to maximize $\log \underline{p}(\mathbf{y}; \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$ makes the approximation as good as possible. The optimal Gaussian density transform $q^*(\boldsymbol{\beta})$ is the $N(\boldsymbol{\mu}^*_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}^*_{q(\boldsymbol{\beta})})$ density function, where $\boldsymbol{\mu}^*_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}^*_{q(\boldsymbol{\beta})}$ are the maximizers of $\log \underline{p}(\mathbf{y}; \boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})$. Newton–Raphson iteration can be used to determine $\boldsymbol{\mu}^*_{q(\boldsymbol{\beta})}$ and $\boldsymbol{\Sigma}^*_{q(\boldsymbol{\beta})}$. Further details may be found in the work of Ormerod (2008).

## 3. TANGENT TRANSFORM APPROACH

Not all variational approximations fit within the Kullback–Leibler divergence framework. Another variety are what might be called *tangent transform* variational approximations since they work with ‘tangent-type’ representations of concave and convex functions. An example of such a representation is

$$\log(x) = \min_{\xi>0}\{\xi x - \log(\xi) - 1\} \quad \text{for all } x > 0. \qquad (20)$$

Figure 7 provides a graphical description of (20).

The representation (20) implies that

$$\log(x) \leq \xi x - \log(\xi) - 1 \quad \text{for all } \xi > 0.$$

The fact that $\xi x - \log(\xi) - 1$ is linear in $x$ for every value of the *variational parameter* $\xi > 0$ allows for simplifications of expressions involving the logarithmic function. The value of $\xi$ can then be chosen to make the approximation as accurate as possible.

Tangent transform variational approximations are underpinned by the theory of *convex duality* (e.g., Rockafellar 1972). We will not delve into that here, and instead stay on course with statistical examples. The interested reader should consult the article by Jordan et al. (1999).

### 3.1 Bayesian Logistic Regression

As described by Jaakkola and Jordan (2000), Bayesian logistic regression lends itself to tangent transform variational approximation. Hence, we consider the Bayesian logistic regression model

$$\begin{aligned} Y_i|\beta_0, \ldots, \beta_k \overset{\text{ind.}}{\sim} \text{Bernoulli}\big([1 + \exp\{-(\beta_0 + \beta_1 x_{1i} + \cdots &\\ + \beta_k x_{ki})\}]^{-1}\big), \qquad 1 \leq i \leq n, \end{aligned}$$

where the prior distribution on the coefficient vector $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_k)$ takes the form $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}})$. The likelihood is

$$p(\mathbf{y}|\boldsymbol{\beta}) = \exp\left[\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T \log\{\mathbf{1}_n + \exp(\mathbf{X}\boldsymbol{\beta})\}\right],$$

where $\mathbf{X} = [1 \;\; x_{1i} \;\cdots\; x_{ki}]_{1\leq i\leq n}$. The posterior density of $\boldsymbol{\beta}$ is

$$p(\boldsymbol{\beta}|\mathbf{y}) = p(\mathbf{y}, \boldsymbol{\beta}) \Big/ \int_{\mathbb{R}^{k+1}} p(\mathbf{y}, \boldsymbol{\beta})\, d\boldsymbol{\beta},$$

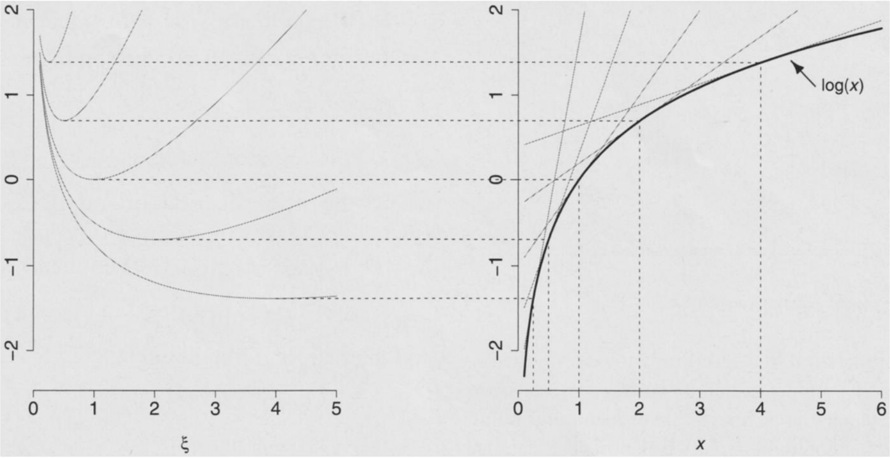

Figure 7. Variational representation of the logarithmic function. *Left axes*: Members of family of functions $f(x, \xi) \equiv \xi x - \log(\xi) - 1$ versus $\xi > 0$, for $x \in \{0.25, 0.5, 1, 2, 4\}$, shown as gray curves. *Right axes*: For each $x$, the minimum of $f(x, \xi)$ over $\xi$ corresponds to $\log(x)$. In the $x$ direction the $f(x, \xi)$ are linear and are shown in gray.

where

$$p(\mathbf{y},\boldsymbol{\beta}) = \exp\Big[\mathbf{y}^T\mathbf{X}\boldsymbol{\beta} - \mathbf{1}_n^T \log\{\mathbf{1}_n + \exp(\mathbf{X}\boldsymbol{\beta})\} - \frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}})^T \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_{\boldsymbol{\beta}}) - \frac{k+1}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\Big]. \tag{21}$$

Once again, we are stuck with a multivariate intractable integral in the normalizing factor. We get around this by noting the following representation of $-\log(1+e^x)$ as the maxima of a family of parabolas:

$$-\log(1+e^x) = \max_{\xi\in\mathbb{R}}\left\{A(\xi)x^2 - \frac{1}{2}x + C(\xi)\right\} \quad \text{for all } x \in \mathbb{R}, \tag{22}$$

where

$$A(\xi) \equiv -\tanh(\xi/2)/(4\xi) \qquad \text{and}$$
$$C(\xi) \equiv \xi/2 - \log(1+e^{\xi}) + \xi\tanh(\xi/2)/4.$$

While the genesis of (22) may be found in the article by Jaakkola and Jordan (2000), it is easily checked via elementary calculus methods. It follows from (22) that

$$\begin{aligned} -\mathbf{1}_n^T &\log\{\mathbf{1}_n + \exp(\mathbf{X}\boldsymbol{\beta})\} \\ &\geq \mathbf{1}_n^T\left\{A(\boldsymbol{\xi})\odot(\mathbf{X}\boldsymbol{\beta})^2 - \frac{1}{2}\mathbf{X}\boldsymbol{\beta} + C(\boldsymbol{\xi})\right\} \\ &= \boldsymbol{\beta}^T\mathbf{X}^T\operatorname{diag}\{A(\boldsymbol{\xi})\}\mathbf{X}\boldsymbol{\beta} - \frac{1}{2}\mathbf{1}_n^T\mathbf{X}\boldsymbol{\beta} + \mathbf{1}_n^T C(\boldsymbol{\xi}), \end{aligned} \tag{23}$$

where $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)$ is an $n\times 1$ vector of variational parameters. This gives us the following lower bound on $p(\mathbf{y},\boldsymbol{\beta})$:

$$\begin{aligned} \underline{p}(\mathbf{y},\boldsymbol{\beta};\boldsymbol{\xi}) = \exp\bigg(&-\frac{1}{2}\boldsymbol{\beta}^T\big[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} - 2\mathbf{X}^T\operatorname{diag}\{A(\boldsymbol{\xi})\}\mathbf{X}\big]\boldsymbol{\beta} \\ &+ \left\{\left(\mathbf{y}-\frac{1}{2}\mathbf{1}_n\right)^T\mathbf{X} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\right\}\boldsymbol{\beta} \\ &- \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \mathbf{1}_n^T C(\boldsymbol{\xi}) \\ &- \frac{k+1}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}|\bigg) \end{aligned}$$

which is proportional to a Multivariate Normal density in $\boldsymbol{\beta}$. Upon normalization we obtain the following family of variational approximations to $\boldsymbol{\beta}|\mathbf{y}$:

$$\boldsymbol{\beta}|\mathbf{y};\boldsymbol{\xi} \sim N(\boldsymbol{\mu}(\boldsymbol{\xi}),\boldsymbol{\Sigma}(\boldsymbol{\xi})), \tag{24}$$

where

$$\boldsymbol{\Sigma}(\boldsymbol{\xi}) \equiv \big[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} - 2\mathbf{X}^{\mathsf{T}}\operatorname{diag}\{A(\boldsymbol{\xi})\}\mathbf{X}\big]^{-1} \qquad \text{and}$$
$$\boldsymbol{\mu}(\boldsymbol{\xi}) \equiv \boldsymbol{\Sigma}(\boldsymbol{\xi})\left\{\mathbf{X}^{\mathsf{T}}\left(\mathbf{y}-\frac{1}{2}\mathbf{1}\right) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right\}.$$

We are left with the problem of determining the vector of variational parameters $\boldsymbol{\xi}\in\mathbb{R}^n$. A natural way of choosing these is to make

$$\underline{p}(\mathbf{y};\boldsymbol{\xi}) \equiv \int \underline{p}(\mathbf{y},\boldsymbol{\beta};\boldsymbol{\xi})\,d\boldsymbol{\beta}$$

as close as possible to $p(\mathbf{y})$. Since $\underline{p}(\mathbf{y};\boldsymbol{\xi}) \leq p(\mathbf{y})$ for all $\boldsymbol{\xi}$, this reduces to the problem of maximizing $\underline{p}(\mathbf{y};\boldsymbol{\xi})$ over $\boldsymbol{\xi}$. Note that this lower bound on $\log p(\mathbf{y})$ has explicit expression:

$$\begin{aligned} \log\underline{p}(\mathbf{y};\boldsymbol{\xi}) = \frac{1}{2}&\log|\boldsymbol{\Sigma}(\boldsymbol{\xi})| - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \\ &+ \frac{1}{2}\boldsymbol{\mu}(\boldsymbol{\xi})^T\boldsymbol{\Sigma}(\boldsymbol{\xi})^{-1}\boldsymbol{\mu}(\boldsymbol{\xi}) - \frac{1}{2}\boldsymbol{\mu}_{\boldsymbol{\beta}}^T\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ &+ \sum_{i=1}^n\{\xi_i/2 - \log(1+e^{\xi_i}) + (\xi_i/4)\tanh(\xi_i/2)\}. \end{aligned}$$

Even though this can be maximized numerically in a similar fashion to (19), Jaakkola and Jordan (2000) derived a simpler algorithm based on the notion of Expectation Maximization (EM) (e.g., McLachlan and Krishnan 1997) with $\boldsymbol{\beta}$ playing the role of a set of latent variables. Treating $\mathbf{y},\boldsymbol{\beta}$ as the set of 'complete data,' the E-step of their EM algorithm involves

$$Q(\boldsymbol{\xi}^{\text{new}}|\boldsymbol{\xi}) \equiv E_{\boldsymbol{\beta}|\mathbf{y};\boldsymbol{\xi}}\{\log\underline{p}(\mathbf{y},\boldsymbol{\beta};\boldsymbol{\xi}^{\text{new}})\},$$

where $\underline{p}(\mathbf{y},\boldsymbol{\beta};\boldsymbol{\xi})$ is interpreted as the variational lower bound on the 'complete data likelihood.' This results in the explicit expression

$$\begin{aligned} Q(\boldsymbol{\xi}^{\text{new}}|\boldsymbol{\xi}) &= \operatorname{tr}\big[\mathbf{X}^T\operatorname{diag}\{A(\boldsymbol{\xi}^{\text{new}})\}\mathbf{X}\{\boldsymbol{\Sigma}(\boldsymbol{\xi}) + \boldsymbol{\mu}(\boldsymbol{\xi})\boldsymbol{\mu}(\boldsymbol{\xi})^T\}\big] \\ &\quad + \mathbf{1}_n^T C(\boldsymbol{\xi}^{\text{new}}) + \text{terms not involving } \boldsymbol{\xi}^{\text{new}}. \end{aligned}$$

Differentiating with respect to $\boldsymbol{\xi}^{\text{new}}$ and using the fact that $A(\boldsymbol{\xi})$ is monotonically increasing over $\xi > 0$, the M-step can be shown to have the exact solution

$$(\boldsymbol{\xi}^{\text{new}})^2 = \operatorname{diagonal}\big[\mathbf{X}\{\boldsymbol{\Sigma}(\boldsymbol{\xi}) + \boldsymbol{\mu}(\boldsymbol{\xi})\boldsymbol{\mu}(\boldsymbol{\xi})^{\mathsf{T}}\}\mathbf{X}^{\mathsf{T}}\big]. \tag{25}$$

Taking positive square roots on both sides of (25) leads to Algorithm 6.

Convergence of Algorithm 6 is monotone and usually quite rapid (Jaakkola and Jordan 2000).

## 4. FREQUENTIST INFERENCE

Up until now, we have only dealt with approximate inference in Bayesian models via variational methods. In this section we point out that variational approximations can be used in frequentist contexts. However, use of variational approximations for frequentist inferential problems is much rarer. Frequentist

**Algorithm 6** Iterative scheme for obtaining the optimal model and variational parameters in the Bayesian logistic regression example.

```
Initialize: $\boldsymbol{\xi}$ ($n \times 1$; all entries positive).
Cycle:
    $\boldsymbol{\Sigma}(\boldsymbol{\xi}) \leftarrow \big[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} - 2\mathbf{X}^{\mathsf{T}}\operatorname{diag}\{A(\boldsymbol{\xi})\}\mathbf{X}\big]^{-1}$,
    $\boldsymbol{\mu}(\boldsymbol{\xi}) \leftarrow \boldsymbol{\Sigma}(\boldsymbol{\xi})\left\{\mathbf{X}^{\mathsf{T}}\left(\mathbf{y}-\frac{1}{2}\mathbf{1}_n\right) + \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}\right\}$,
    $\boldsymbol{\xi} \leftarrow \sqrt{\operatorname{diagonal}\big[\mathbf{X}\{\boldsymbol{\Sigma}(\boldsymbol{\xi}) + \boldsymbol{\mu}(\boldsymbol{\xi})\boldsymbol{\mu}(\boldsymbol{\xi})^{\mathsf{T}}\}\mathbf{X}^{\mathsf{T}}\big]}$
until the increase in $\underline{p}(\mathbf{y};\boldsymbol{\xi})$ is negligible.
```

models that stand to benefit from variational approximations are those for which specification of the likelihood involves conditioning on a vector of latent variables $\mathbf{u}$. In this case, the log-likelihood of the model parameter vector $\boldsymbol{\theta}$ takes the form

$$\ell(\boldsymbol{\theta}) \equiv \log p(\mathbf{y};\boldsymbol{\theta}) = \log \int p(\mathbf{y}|\mathbf{u};\boldsymbol{\theta})p(\mathbf{u};\boldsymbol{\theta})\,d\mathbf{u}. \tag{26}$$

The maximum likelihood estimate of $\boldsymbol{\theta}$ is exactly

$$\widehat{\theta} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\,\ell(\boldsymbol{\theta})$$

but, because of the integral in (26), $\ell(\boldsymbol{\theta})$ may not be available in closed form. Depending on the forms of $p(\mathbf{y}|\mathbf{u};\boldsymbol{\theta})$ and $p(\mathbf{u};\boldsymbol{\theta})$, either the density transform or tangent transform approaches can result in more tractable approximations to $\ell(\boldsymbol{\theta})$. For the remainder of this section we restrict discussion to the density transform approach. The tangent transform approach has a similar treatment.

Let $q(\mathbf{u})$ be an arbitrary density function in $\mathbf{u}$. Repeating the steps given at (2), but with the log marginal likelihood $\log p(\mathbf{y})$ replaced by the log-likelihood $\ell(\boldsymbol{\theta})$, we obtain

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \int q(\mathbf{u})\log\left\{\frac{p(\mathbf{y},\mathbf{u};\boldsymbol{\theta})}{q(\mathbf{u})}\right\}d\mathbf{u}\\ &\quad + \int q(\mathbf{u})\log\left\{\frac{q(\mathbf{u})}{p(\mathbf{u}|\mathbf{y};\boldsymbol{\theta})}\right\}d\mathbf{u}\\ &\geq \underline{\ell}(\boldsymbol{\theta};q),\end{aligned}$$

where

$$\underline{\ell}(\boldsymbol{\theta};q) \equiv \int q(\mathbf{u})\log\left\{\frac{p(\mathbf{y},\mathbf{u};\boldsymbol{\theta})}{q(\mathbf{u})}\right\}d\mathbf{u}. \tag{27}$$

We now have the option of choosing $q$ to make $\underline{\ell}(\boldsymbol{\theta};q)$ more tractable while also aiming to minimize the Kullback–Leibler divergence between $q$ and $p(\mathbf{u}|\mathbf{y};\boldsymbol{\theta})$. In theory, the product density methodology of Section 2.2 could be used to guide the choice of $q$. However, we have yet to find a nontrivial frequentist example where an explicit solution arises. Suppose, instead, that we restrict $q$ to a parametric family of densities $\{q(\mathbf{u};\boldsymbol{\xi}):\boldsymbol{\xi}\in\Xi\}$. Then the log-likelihood lower bound (27) becomes

$$\underline{\ell}(\boldsymbol{\theta},\boldsymbol{\xi};q) = \int q(\mathbf{u};\boldsymbol{\xi})\log\left\{\frac{p(\mathbf{y},\mathbf{u};\boldsymbol{\theta})}{q(\mathbf{u};\boldsymbol{\xi})}\right\}d\mathbf{u}. \tag{28}$$

We should maximize over the *variational parameters* $\boldsymbol{\xi}$ to minimize the Kullback–Leibler divergence between $q(\mathbf{u};\boldsymbol{\xi})$ and $p(\mathbf{u}|\mathbf{y};\boldsymbol{\theta})$, and over the *model parameters* $\boldsymbol{\theta}$ to maximize the approximate log-likelihood. This leads to the new maximization problem:

$$(\widehat{\underline{\boldsymbol{\theta}}},\widehat{\underline{\boldsymbol{\xi}}}) = \underset{\boldsymbol{\theta},\boldsymbol{\xi}}{\operatorname{argmax}}\,\underline{\ell}(\boldsymbol{\theta},\boldsymbol{\xi};q).$$

Then $\widehat{\underline{\boldsymbol{\theta}}}$ is a variational approximation to the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$. Standard error estimates can be obtained by plugging in $\widehat{\underline{\boldsymbol{\theta}}}$ for $\boldsymbol{\theta}$ and $\underline{\boldsymbol{\xi}}$ for $\boldsymbol{\xi}$ in the variational approximate Fisher information matrix, the matrix that arises from replacement of $\ell(\boldsymbol{\theta})$ by $\underline{\ell}(\boldsymbol{\theta},\boldsymbol{\xi};q)$ in the definition of Fisher information. However, to our knowledge, asymptotic normality theory that justifies such standard error estimation has not yet been done.

### 4.1 Poisson Mixed Model

Consider the (non-Bayesian) Poisson mixed model

$$\begin{aligned} Y_{ij}|U_i &\overset{\text{ind.}}{\sim} \text{Poisson}\{\exp(\beta_0+\beta_1 x_{ij}+U_i)\},\\ U_i &\overset{\text{ind.}}{\sim} N(0,\sigma^2), \qquad 1\leq j\leq n_i, 1\leq i\leq m,\end{aligned} \tag{29}$$

where $y_{ij}$ is the $j$th response measurement for unit $i$, and the deterministic predictors $x_{ij}$ are defined similarly. The log-likelihood of $(\beta_0,\beta_1,\sigma^2)$ involves intractable integrals, but the lower bound (28) takes the form

$$\begin{aligned}&\underline{\ell}(\beta_0,\beta_1,\sigma^2;q)\\ &= \int_{\mathbb{R}^m}\left(\sum_{i=1}^m\left[\sum_{j=1}^{n_i}\{y_{ij}(\beta_0+\beta_1 x_{ij}+u_i)\right.\right.\\ &\qquad \left. - e^{\beta_0+\beta_1 x_{ij}+u_i} - \log(y_{ij}!)\} - \frac{u_i^2}{2\sigma^2}\right]\\ &\qquad \left. - \frac{m}{2}\log(2\pi\sigma^2) - \log q(u_1,\ldots,u_m)\right)\\ &\qquad \times q(u_1,\ldots,u_m)\,du_1\cdots du_m.\end{aligned}$$

Setting $q$ to be the product of $m$ univariate Normal densities with mean $\mu_i$ and variance $\lambda_i > 0$, $1\leq i\leq m$, leads to the closed form lower bound:

$$\begin{aligned}&\underline{\ell}(\beta_0,\beta_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda};q)\\ &= \sum_{i=1}^m\sum_{j=1}^{n_i}\{y_{ij}(\beta_0+\beta_1 x_{ij}+\mu_i)\\ &\qquad + e^{\beta_0+\beta_1 x_{ij}+\mu_i+(1/2)\lambda_i} - \log(y_{ij}!)\}\\ &\qquad + \frac{m}{2}\{1-\log(\sigma^2)\} + \frac{1}{2}\sum_{i=1}^m\left\{\log(\lambda_i) - \frac{\mu_i^2+\lambda_i}{\sigma^2}\right\}\end{aligned}$$

for all values of the variational parameters $\boldsymbol{\mu} = (\mu_1,\ldots,\mu_m)$ and $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_m)$. Maximizing over these parameters narrows the gap between $\underline{\ell}(\beta_0,\beta_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda};q)$ and $\ell(\beta_0,\beta_1,\sigma^2)$ and so sensible estimators of the model parameters are

$$(\widehat{\underline{\beta_0}},\widehat{\underline{\beta_1}},\widehat{\underline{\sigma}}^2) = (\beta_0,\beta_1,\sigma^2)$$

component of

$$\underset{\beta_0,\beta_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda}}{\operatorname{argmax}}\ \underline{\ell}(\beta_0,\beta_1,\sigma^2,\boldsymbol{\mu},\boldsymbol{\lambda};q).$$

Recently, Hall, Ormerod, and Wand (2010) established consistency and rates of convergence results for $\widehat{\underline{\beta_0}}$, $\widehat{\underline{\beta_1}}$, and $\widehat{\underline{\sigma}}^2$.

## 5. CLOSING DISCUSSION

Our goal in this article is to explain variational approximations in a digestible form for a statistical audience. As mentioned in the Introduction, the important issue of *accuracy* of variational approximations is not dealt with here. The expositions by Jordan (2004) and Titterington (2004) provide access to some of the literature on variational approximation accuracy.

Variational approximations have the potential to become an important player in statistical inference. New variational approximation methods are continually being developed. The recent emergence of formal software for variational inference is certain to accelerate its widespread use. The usefulness of variational approximations increases as the size of the problem increases and Monte Carlo methods such as MCMC start to become untenable.

*[Received March 2009. Revised April 2010.]*

## REFERENCES

Albert, J. H., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679. [146]

Archambeau, C., Cornford, D., Opper, M., and Shawe-Taylor, J. (2007), "Gaussian Process Approximations of Stochastic Differential Equations," *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 1, 1–16. [149]

Barber, D., and Bishop, C. M. (1998), "Ensemble Learning for Multi-Layer Networks," in *Advances in Neural Information Processing Systems 10*, eds. M. I. Jordan, K. J. Kearns, and S. A. Solla, Cambridge, MA: MIT Press, pp. 395–401. [149]

Bishop, C. M. (2006), *Pattern Recognition and Machine Learning*, New York: Springer. [140,145,148]

Boyd, S., and Vandenberghe, L. (2004), *Convex Optimization*, Cambridge: Cambridge University Press. [143]

Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174. [143]

Consonni, G., and Marin, J.-M. (2007), "Mean-Field Variational Approximate Bayesian Inference for Latent Variable Models," *Computational Statistics and Data Analysis*, 52, 790–798. [146]

Girolami, M., and Rogers, S. (2006), "Variational Bayesian Multinomial Probit Regression," *Neural Computation*, 18, 1790–1817. [146]

Hall, P., Humphreys, K., and Titterington, D. M. (2002), "On the Adequacy of Variational Lower Bound Functions for Likelihood-Based Inference in Markovian Models With Missing Values," *Journal of the Royal Statistical Society, Ser. B*, 64, 549–564. [140]

Hall, P., Ormerod, J. T., and Wand, M. P. (2010), "Theory of Gaussian Variational Approximation for a Poisson Linear Mixed Model," *Statistica Sinica*, to appear. [152]

Honkela, A., and Valpola, H. (2005), "Unsupervised Variational Bayesian Learning of Nonlinear Models," in *Advances in Neural Information Processing Systems 17*, eds. L. K. Saul, Y. Weiss, and L. Bottou, Cambridge, MA: MIT Press, pp. 593–600. [149]

Jaakkola, T. S., and Jordan, M. I. (2000), "Bayesian Parameter Estimation via Variational Methods," *Statistics and Computing*, 10, 25–37. [150,151]

Jordan, M. I. (2004), "Graphical Models," *Statistical Science*, 19, 140–155. [140,152]

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999), "An Introduction to Variational Methods for Graphical Models," *Machine Learning*, 37, 183–233. [140,150]

Kass, R. E., and Raftery, A. E. (1995), "Bayes Factors and Model Uncertainty," *Journal of the American Statistical Association*, 90, 773–795. [142]

Kullback, S., and Leibler, R. A. (1951), "On Information and Sufficiency," *The Annals of Mathematical Statistics*, 22, 79–86. [142]

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008), *Generalized, Linear, and Mixed Models* (2nd ed.), New York: Wiley. [144]

McGrory, C. A., and Titterington, D. M. (2007), "Variational Approximations in Bayesian Model Selection for Finite Mixture Distributions," *Computational Statistics and Data Analysis*, 51, 5352–5367. [140,148]

McGrory, C. A., Titterington, D. M., Reeves, R., and Pettitt, A. N. (2009), "Variational Bayes for Estimating the Parameters of a Hidden Potts Model," *Statistics and Computing*, 19, 329–340. [140]

McLachlan, G. J., and Krishnan, T. (1997), *The EM Algorithm and Extensions*, New York: Wiley-Interscience. [151]

Minka, T., Winn, J., Guiver, G., and Kannan, A. (2009), *Infer.Net 2.3*, Cambridge, U.K.: Microsoft Research Cambridge. [140]

Ormerod, J. T. (2008), "On Semiparametric Regression and Data Mining," Ph.D. thesis, School of Mathematics and Statistics, The University of New South Wales, Sydney, Australia. [150]

Parisi, G. (1988), *Statistical Field Theory*, Redwood City, CA: Addison-Wesley. [142]

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, San Mateo, CA: Morgan Kaufmann. [143,145]

Pinheiro, J. C., and Bates, D. M. (2000), *Mixed-Effects Models in S and S-PLUS*, New York: Springer. [146]

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and the R Core Team (2009), "nlme: Linear and Nonlinear Mixed Effects Models," R package version 3.1-93. [146]

R Development Core Team (2010), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. Available at *http://www.R-project.org*. [146]

Rockafellar, R. (1972), *Convex Analysis*, Princeton: Princeton University Press. [150]

Seeger, M. (2000), "Bayesian Model Selection for Support Vector Machines, Gaussian Processes and Other Kernel Classifiers," in *Advances in Neural Information Processing Systems 12*, eds. S. A. Solla, T. K. Leen, and K.-R. Müller, Cambridge, MA: MIT Press, pp. 603–609. [149]

——— (2004), "Gaussian Processes for Machine Learning," *International Journal of Neural Systems*, 14, 69–106. [149]

Teschendorff, A. E., Wang, Y., Barbosa-Morais, N. L., Brenton, J. D., and Caldas, C. (2005), "A Variational Bayesian Mixture Modelling Framework for Cluster Analysis of Gene-Expression Data," *Bioinformatics*, 21, 3025–3033. [140]

Titterington, D. M. (2004), "Bayesian Methods for Neural Networks and Related Models," *Statistical Science*, 19, 128–139. [140,142,152]

Venables, W. N., and Ripley, B. D. (2009), "MASS: Functions and Datasets to Support Venables and Ripley, 'Modern Applied Statistics With S' (4th ed.)," R package version 7.2-48. [148]

Wang, B., and Titterington, D. M. (2006), "Convergence Properties of a General Algorithm for Calculating Variational Bayesian Estimates for a Normal Mixture Model," *Bayesian Analysis*, 1, 625–650. [140]

Winn, J., and Bishop, C. M. (2005), "Variational Message Passing," *Journal of Machine Learning Research*, 6, 661–694. [143]