Published in final edited form as:

J Comput Graph Stat. 2016 ; 25(1): 225–245. doi:10.1080/10618600.2014.983642.

Laplace Variational Approximation for Semiparametric Regression in the Presence of Heteroskedastic Errors

Bruce D. Bugbee [Postdoctoral Fellow]

Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, 77030 (bdbugbee@mdanderson.org)

F. Jay Breidt [Professor]

Department of Statistics, Colorado State University, Fort Collins, CO, 80523 (jbreidt@stat.colostate.edu)

Mark J. van der Woerd [Research Scientist]

Department of Biochemistry & Molecular Biology, Colorado State University, Fort Collins, CO, 80523 (Mark.van der Woerd@colostate.edu) October 12, 2014

Abstract

Variational approximations provide fast, deterministic alternatives to Markov Chain Monte Carlo for Bayesian inference on the parameters of complex, hierarchical models. Variational approximations are often limited in practicality in the absence of conjugate posterior distributions. Recent work has focused on the application of variational methods to models with only partial conjugacy, such as in semiparametric regression with heteroskedastic errors. Here, both the mean and log variance functions are modeled as smooth functions of covariates. For this problem, we derive a mean field variational approximation with an embedded Laplace approximation to account for the non-conjugate structure. Empirical results with simulated and real data show that our approximate method has significant computational advantages over traditional Markov Chain Monte Carlo; in this case, a delayed rejection adaptive Metropolis algorithm. The variational approximation is much faster and eliminates the need for tuning parameter selection, achieves good fits for both the mean and log variance functions, and reasonably reflects the posterior uncertainty. We apply the methods to log-intensity data from a small angle X-ray scattering experiment, in which properly accounting for the smooth heteroskedasticity leads to significant improvements in posterior inference for key physical characteristics of an organic molecule.

Keywords

mean field variational Bayes; penalized splines; small angle X-ray scattering

Supplemental Materials: Variational Method Code. R code for performing the variational approximation method described in this paper is included. Also with this code are instructions for creating the simulated data sets described here as well as the experimental SAXS data used. (Zip file)

1 Introduction

Variational approximations allow for the approximate computation of posterior parameter distributions of complex models in a deterministic fashion (Wainwright and Jordan, 2008). The term *mean field variational Bayes* (MFVB) refers to a class of variational approximations to posterior distributions under a nonparametric product density constraint (Bishop, 2006; Ormerod and Wand, 2010). Applications of MFVB approximations include mixtures (Attias, 2000), linear mixed models (Ormerod and Wand, 2010), and nonparametric regression with missing data (Faes et al., 2011). Given the connection between MFVB and full posterior parameter conditionals, variational approximations can be directly implemented in situations where all parameters have known posterior conditionals and Gibbs sampling is appropriate (Casella and George, 1992). This has led to useful software packages such as Infer.NET that allow for “black box” computation of variational approximations given fully known posterior conditionals (Minka et al., 2010; Wang and Wand, 2011).

Recent studies have considered *non-conjugate models* where one or more of the posterior conditionals are not of known form (Carlin and Polson, 1991). Examples of such models include logistic regression (Jaakkola and Jordan, 1997), nonparametric regression with measurement error (Pham et al., 2013), generalized linear mixed models (Tan and Nott, 2013), and correlated topic models (Blei and Lafferty, 2007). Variational approximations of non-conjugate models often have to be derived on a case by case basis and can be difficult to handle. Wand et al. (2011) highlight the use of auxiliary variables, univariate quadrature methods, and finite mixture approximations for addressing hierarchical models with complex distributions. Knowles and Minka (2011) focus on generalizing non-conjugate variational inference for exponential family cases arising from multinomial and binary regression. Salimans and Knowles (2013) and Nott et al. (2013) have recently developed alternatives to standard mean field variational approximations, called structural or fixed-form variational Bayes. These methods have shown promise for addressing non-conjugate problems through the assumption of exponential forms and application of Monte Carlo integration.

One such non-conjugate model arises from semiparametric regression with heteroskedastic errors. Models of this type have been explored in Ruppert et al. (2003), Chan et al. (2006), and Crainiceanu et al. (2007). Using the mixed model formulation of penalized spline regression described in Ruppert et al. (2003), models of this type allow for parametric and nonparametric terms in both the mean and variance levels. Nott et al. (2012) look at the case of a linear fixed effect model with heteroskedastic errors. The work we present here is a novel application of variational approximation methods to the case of semiparametric regression in the presence of heteroskedastic errors, represented as multilevel linear mixed effect models. We use an embedded Laplace approximation to handle the non-conjugacy found in our variational approximation (Wang and Blei, 2013; Nott et al., 2012). Results of Salimans and Knowles (2013) and Nott et al. (2013) could also be applied in this context, though our method uses analytic formulas for the variational moments needed in our special case rather than relying on Monte Carlo integration.

Our work is motivated by data arising from small angle X-ray scattering (SAXS), an imaging technique used in the estimation of physical shape characteristics of complex macromolecules. The general experimental methodology consists of bombarding a sample of the molecule in solution with a beam of X-rays and measuring the resulting scattering pattern on a two-dimensional detector orthogonal to the X-ray beam. The scattering data are reduced from two dimensions to one by averaging intensity (photon counts on the detector) across all of the pixels at a given radial distance from a fixed center. The radii in the detector plane are equivalent to scattering angles relative to the incident X-ray beam: zero radius corresponds to the beam itself, and larger angles correspond to increasing radii. The resulting “intensity curve” is then the set of ordered pairs (angle, radial average intensity at that angle). See Figure 4. The one-dimensional data from SAXS experiments are heteroskedastic, due to characteristics of the molecule in solution and the measurement mechanism; see for example Breidt et al. (2012). SAXS is a low-resolution method in which physical parameters such as maximum linear particle dimension and squared radius of gyration (R_g^2) can be estimated. The squared radius of gyration of a particle is defined as the ratio of the particle’s moment of inertia and mass. Appropriate interval estimates for such parameters are of great importance to experimental scientists, who develop theoretical models consistent with the experimental scattering data. Accounting for the heteroskedasticity in the SAXS data is essential for proper inference on these physical parameters, which in turn allows experimental scientists to accurately determine the class of molecular structures that is consistent with a given SAXS dataset.

Fast analytic methods are desirable for SAXS data. Exposure times on the order of one second produce a single SAXS image and data set, and experiments generating many images across a range of experimental conditions are common. We will demonstrate that our variational/Laplace method is substantially faster than a Markov Chain Monte Carlo alternative.

Unless otherwise stated, all random vectors $\mathbf{z}$ are assumed to be continuous with densities of the form $p(\mathbf{z})$. Conditional probability densities will be written as $p(\mathbf{z} | \mathbf{w})$ with $p(\mathbf{z} | \cdot)$ representing the density conditional on all other random quantities in the model. Let $(x - \kappa)_+^p = \max\{0, (x - \kappa)^p\}$. Diagonal matrices are written as $\text{diag}(a_1, \dots, a_m)$ and block diagonal matrices as $\text{blockdiag}(A_1, \dots, A_m)$. Identity matrices of dimension p are written as $\mathcal{I}_p$. Finally, let $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ be a p -dimensional multivariate Normal distribution and $\mathcal{IG}(A, B)$ be an Inverse Gamma distribution with shape and scale parameters A and B .

2 Variational Bayes

We first provide a brief background on the theory of variational approximations. Let $\mathbf{y}$ be a response vector and $\boldsymbol{\psi}$ be a vector of parameters. Direct evaluation of $p(\mathbf{y})$ in the posterior parameter distribution

$$p(\psi|\mathbf{y}) = \frac{p(\mathbf{y}, \psi)}{p(\mathbf{y})} \quad (1)$$

generally requires high dimensional integration, which may not be analytically tractable. A standard approach is to use Markov Chain Monte Carlo (MCMC) methods to approximate $p(\boldsymbol{\psi}|\mathbf{y})$. However, as models grow in complexity, the computational resources needed to perform MCMC increase dramatically and thus fast approximations become valuable.

Let $q(\boldsymbol{\psi})$ be an arbitrary density function over the parameter space Ψ . Then

$$\begin{aligned} \log p(\mathbf{y}) &= \int_{\Psi} q(\psi) \log p(\mathbf{y}) d\psi \\ &\geq \int_{\Psi} q(\psi) \log \left\{ \frac{p(\mathbf{y}, \psi)}{q(\psi)} \right\} d\psi \\ &=: \log p(\mathbf{y}; q), \end{aligned} \quad (2)$$

where the inequality follows from Kullback and Leibler (1951), who show the Kullback-Liebler (K-L) divergence $\int q(\psi) \log \left\{ \frac{q(\psi)}{p(\psi|\mathbf{y})} \right\} d\psi \geq 0$ with equality if and only if $q(\boldsymbol{\psi}) = p(\boldsymbol{\psi}|\mathbf{y})$. The quantity $\log p(\mathbf{y}; q)$ is commonly known as the *evidence lower bound*. The closer our choice of $q(\boldsymbol{\psi})$ is to the true posterior distribution, the smaller the gap between $p(\mathbf{y})$ and $p(\mathbf{y}; q)$. Minimizing the gap between these two quantities is equivalent to minimizing the K-L divergence. Variational approximations are useful when analysis of $p(\mathbf{y}; q)$ is easier than $p(\mathbf{y})$, which can be achieved by restricting $q(\boldsymbol{\psi})$ in some manner.

2.1 Defining $q(\boldsymbol{\psi})$

The most common restriction imposed on the density $q(\boldsymbol{\psi})$, which we use here, is that for some partition of the parameter vector $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots, \boldsymbol{\psi}_L)$,

$$q(\psi) = \prod_{l=1}^L q_l(\psi_l). \quad (3)$$

Variational approximations that impose such a product-type density restriction on $q(\boldsymbol{\psi})$ are referred to as *mean field variational approximations*. When this restriction is used in a Bayesian setting, the term *mean field variational Bayes* or *variational Bayes* for short has become standard. Often there is a somewhat natural partition of the parameter vector given the structure of the model of interest. Both Titterton (2004) and Ormerod and Wand (2010) discuss the choice of the partition. Some partitions (e.g. partitions where there is a high degree of posterior dependence between parameters across partitions) can lead to poor approximations. Alternative forms of variational approximations can be achieved by constraining $q(\boldsymbol{\psi})$ to a class of known parametric density functions but will not be discussed here.

Adopting the notation in Ormerod and Wand (2010), the optimal variational densities, $q_1^*(\psi_1), q_2^*(\psi_2), \dots, q_L^*(\psi_L)$, that minimize the K-L divergence described in (2) can be found using the relationship

$$q_l^*(\psi_l) \propto \exp [E_{-\psi_l} \{\log p(\psi_l | \cdot)\}]. \quad (4)$$

Here $p(\boldsymbol{\psi}_l | \cdot)$ is the posterior conditional distribution

$$p(\psi_l | \mathbf{y}, \psi_1, \dots, \psi_{l-1}, \psi_{l+1}, \dots, \psi_L). \quad (5)$$

The expectation $E_{-\boldsymbol{\psi}_l}$ is with respect to all q -densities except for $q(\boldsymbol{\psi}_l)$. The dependence of the optimal q -densities on the posterior conditional distributions indicates a close tie to Gibbs sampling (Casella and George, 1992). In fact, when conjugate priors are used for $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_L$ and the posterior conditionals have known forms, the optimal q -densities will have the same forms but different parameterizations.

Implementing a variational Bayes approximation under these conditions amounts to determining a set of *variational parameters* that fully define each q -density. Once these variational parameters are determined, the estimates of the model parameters $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_L$ are taken from the densities $q_1^*, \dots, q_L^*$, usually as the means associated with each density.

3 Penalized Splines With Heteroskedastic Errors

A simple case of regression models in which both the response and the associated variances are smoothly varying across some underlying set of covariates is

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_i^2), \quad \log(\sigma_i^2) = g(x_i) \quad (i = 1, 2, \dots, N). \quad (6)$$

Adopting a penalized spline formulation as described in Ruppert et al. (2003), the nonparametric model takes the form for all $i = 1, 2, \dots, N$:

$$\begin{aligned} y_i &= \left[1, x_i, \dots, x_i^p, (x_i - \kappa_1)_+^p, \dots, (x_i - \kappa_K)_+^p \right] \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} + \epsilon_i = \mathbf{C}_i^T \begin{bmatrix} \beta \\ \mathbf{b} \end{bmatrix} + \epsilon_i \\ \mathbf{b} &= (b_1, b_2, \dots, b_K)^T \sim \mathcal{N}_K(\mathbf{0}, \sigma_{\mathbf{b}}^2 \mathcal{J}_{\mathbf{K}}) \\ \epsilon_i &\sim \mathcal{N}(0, \sigma_i^2) \\ \log(\sigma_i^2) &= \left[1, x_i, \dots, x_i^r, (x_i - \kappa_1^V)_+^r, \dots, (x_i - \kappa_{K_V}^V)_+^r \right] \begin{bmatrix} \delta \\ \mathbf{c} \end{bmatrix} = \mathbf{C}_{\mathbf{v}i}^T \begin{bmatrix} \delta \\ \mathbf{c} \end{bmatrix} \\ \mathbf{c} &= (c_1, c_2, \dots, c_{K_V})^T \sim \mathcal{N}_{K_V}(\mathbf{0}, \sigma_{\mathbf{c}}^2 \mathcal{J}_{\mathbf{K}_V}), \end{aligned} \quad (7)$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ and $\boldsymbol{\delta} = (\delta_0, \delta_1, \dots, \delta_r)^T$ are unknown parameter vectors and $\{\kappa_k\}_{k=1}^K$ and $\{\kappa_k^V\}_{k=1}^{K_V}$ are sets of fixed, known knots.

Define the $N \times (p + K + 1)$ and $N \times (r + K_V + 1)$ design matrices $\mathbf{C} = [\mathbf{C}_i^T]_{i=1}^N$ and $\mathbf{C}_V = [\mathbf{C}_{V_i}^T]_{i=1}^N$ where the i th row corresponds to the i th vector of covariates from (7).

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}^T, \mathbf{b}^T)$, $\boldsymbol{\theta}_V = (\boldsymbol{\delta}^T, \mathbf{c}^T)$, $\Sigma = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$ and

$\mathbf{V} = (\log(\sigma_1^2), \dots, \log(\sigma_N^2))^T$. It is convenient to view (7) as two mixed models with fixed effects $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ and random effects $\mathbf{b}$ and $\mathbf{c}$. The parameters of interest for the model are $\boldsymbol{\theta}$, $\boldsymbol{\theta}_V$, $\sigma_{\mathbf{b}}^2$ and $\sigma_{\mathbf{c}}^2$.

For simplicity, we present our results in the context of the nonparametric model of (7). Minimal changes are required to handle more complex semiparametric structures. Additional parametric covariates can be included by expanding the fixed effect parameter vector and additional non-parametric components or parametric-by-nonparametric interactions can be handled through the inclusion of new random effects with corresponding variance terms.

The truncated polynomial spline basis is chosen both for illustrative purposes and for its usefulness in estimation of a key molecular property from SAXS data, as described in §6. The lack of orthogonality in this basis can cause computational problems, particularly for MCMC methods. Often, P-splines (Eilers and Marx, 1996) or O'Sullivan splines (Wand and Ormerod, 2008) are used as alternatives for their computational stability. Other common choices of basis functions include radial basis functions, B-splines, and thin plate splines, depending on the application. These basis choices affect the definition of the $\mathbf{X}$ and $\mathbf{Z}$ matrices but not the variational approximation method presented here.

Equation (7) can be considered a submodel of the class of spatially adaptive penalized spline regression with heteroskedastic errors described in Crainiceanu et al. (2007). These models allow for both the error variance (σ^2) and the random effects variance ($\sigma_{\mathbf{b}}^2$) to be smoothly varying over some set of covariates. Our model is a simplification that holds $\sigma_{\mathbf{b}}^2$ as a univariate parameter to be estimated. A limitation of a model of this form is that it can have trouble fitting data in regions where dramatic increases in variation coincide with strong changes in curvature. The inclusion of a spatially adaptive random effect variance allows the model to handle changes in the curvature of the underlying function $f(x)$ better. The work presented here focuses on cases where the underlying first derivative of $f(x)$ appears to vary slowly and does not coincide with sharp increases of variation, as is the case for our motivating application to SAXS data.

3.1 Bayesian Framework

Assume $\beta_{\mathcal{N}_p}(\mathbf{0}, \sigma_{\beta}^2 \mathcal{I}_p)$ and $\delta_{\mathcal{N}_r}(\mathbf{0}, \sigma_{\delta}^2 \mathcal{I}_r)$ where σ_{β}^2 and σ_{δ}^2 are given hyperparameters.

Assume Inverse Gamma priors $\sigma_{\mathbf{b}}^2 \mathcal{IG}(A_{\mathbf{b}}, B_{\mathbf{b}})$ and $\sigma_{\mathbf{c}}^2 \mathcal{IG}(A_{\mathbf{c}}, B_{\mathbf{c}})$ with hyperparameters

$A_{\mathbf{b}}, A_{\mathbf{c}}, B_{\mathbf{b}},$ and $B_{\mathbf{c}}$. Define $\Sigma_{\theta} = \text{blockdiag} \left\{ \sigma_{\beta}^2 \mathcal{I}_p, \sigma_{\mathbf{b}}^2 \mathcal{I}_K \right\}$ and $\Sigma_{\theta_V} = \text{blockdiag} \left\{ \sigma_{\delta}^2 \mathcal{I}_r, \sigma_{\mathbf{c}}^2 \mathcal{I}_{K_V} \right\}$.

As in Crainiceanu et al. (2007), the model structure and conjugate nature of the priors described here lend themselves to a hybrid Gibbs sampling computational approach to fitting our model. The posterior conditionals of the parameters of interest are

$$\begin{aligned} \theta &| \cdot \sim \mathcal{N} \left(\mathbf{M} \mathbf{C}^T \Sigma^{-1} \mathbf{y}, \mathbf{M} \right) \\ \mathbf{M} &= \left(\mathbf{C}^T \Sigma^{-1} \mathbf{C} + \Sigma_{\theta}^{-1} \right)^{-1} \\ \sigma_{\mathbf{b}}^2 &| \cdot \sim \mathcal{IG} \left(A_{\mathbf{b}} + \frac{K}{2}, B_{\mathbf{b}} + \frac{\|\mathbf{b}\|^2}{2} \right) \\ \sigma_{\mathbf{c}}^2 &| \cdot \sim \mathcal{IG} \left(A_{\mathbf{c}} + \frac{K_V}{2}, B_{\mathbf{c}} + \frac{\|\mathbf{c}\|^2}{2} \right) \\ p(\theta_V | \cdot) &\propto \exp \left[-\frac{1}{2} \left\{ \sum_{i=1}^N \mathbf{C}_{V_i}^T \theta_V + \sum_{i=1}^N \left(Y_i - \mathbf{C}_i^T \theta \right)^2 \exp \left(-\mathbf{C}_{V_i}^T \theta_V \right) + \theta_V^T \Sigma_{\theta_V}^{-1} \theta_V \right\} \right]. \end{aligned} \quad (8)$$

The challenge of fitting this model is due to the non-standard distribution of $\theta_V | \cdot$. Crainiceanu et al. (2007) address this problem by introducing a latent error term to the variance model that allows for the sampling from N univariate posterior conditionals, which can be accomplished through traditional Metropolis-Hastings schemes. Another possibility is the use of methods designed for sampling from multivariate parameter distributions, such as the Delayed Rejection Adaptive Metropolis (DRAM) method of Haario et al. (2006).

4 Variational Approximation of Heteroskedastic Model

We now derive an approximation method for fitting the model (7) via variational Bayes. Nott et al. (2012) describe a variational approximation for the standard linear model in the presence of heteroskedasticity of known parametric form. Our approximation applies more broadly, to linear mixed models generally and to semiparametric regression under heteroskedasticity specifically. Unlike Nott et al. (2012), we do not assume a parametric form for the heteroskedasticity, but model it flexibly with penalized splines.

Let $\psi = \left(\theta, \theta_V, \sigma_{\mathbf{b}}^2, \sigma_{\mathbf{c}}^2 \right)^T$ be a vector containing all parameters of interest. Using the product density constraint from (3), we assume that the variational density $q(\psi)$ is of the form

$$q(\psi) = q_1(\theta) q_2(\theta_V) q_3(\sigma_{\mathbf{b}}^2) q_4(\sigma_{\mathbf{c}}^2). \quad (9)$$

For notational simplicity, the subscript index on each component density of $q(\psi)$ will be dropped in future references (e.g. $q_1(\theta) = q(\theta)$). The notation q^* will be used to describe the optimal distributions of a parameter derived using the relationship described in (4).

Define $\mathbf{\Gamma} = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_N)$ where $\gamma_i = E_{-\theta} \left[\exp \left(-\mathbf{C}_{\mathbf{V}_i}^T \theta_V \right) \right]$ corresponds to the moment generating function of θ_V under the q -density evaluated at $-\mathbf{C}_{\mathbf{V}_i}$. Each parameter's q -distribution will be defined by a set of variational parameters, denoted with a subscript

$q(\cdot)$. Let $\mu_{q(1/\sigma_b^2)} = E_{-\theta} \left[1/\sigma_b^2 \right]$, $\mu_{q(1/\sigma_c^2)} = E_{-\theta_V} \left[1/\sigma_c^2 \right]$,

$$\Sigma_{q(\theta)} = \left\{ \text{blockdiag} \left(\frac{1}{\sigma_\beta^2} \mathcal{J}_p, \mu_{q(1/\sigma_b^2)} \mathcal{J}_K \right) + \mathbf{C}^T \mathbf{\Gamma} \mathbf{C} \right\}^{-1}, \text{ and } \mu_{q(\theta)} = \Sigma_{q(\theta)} \mathbf{C}^T \mathbf{\Gamma} \mathbf{y}.$$

Using the relationship found in (4) and the posterior conditionals described in (8), the optimal q -distributions for our model are

$$\begin{aligned} \theta &\stackrel{q^*}{\sim} \mathcal{N} \left(\mu_{q(\theta)}, \Sigma_{q(\theta)} \right) \\ \sigma_b^2 &\stackrel{q^*}{\sim} \mathcal{JG} \left(A_b + \frac{K}{2}, B_b + \frac{1}{2} \parallel \mu_{q(b)} \parallel^2 + \frac{1}{2} \text{trace} \left(\Sigma_{q(b)} \right) \right) \\ \sigma_c^2 &\stackrel{q^*}{\sim} \mathcal{JG} \left(A_c + \frac{K_V}{2}, B_c + \frac{1}{2} \parallel \mu_{q(c)} \parallel^2 + \frac{1}{2} \text{trace} \left(\Sigma_{q(c)} \right) \right) \\ q^*(\theta_V) &\propto \exp \left[-\frac{1}{2} \left\{ \sum_{i=1}^N \mathbf{C}_{\mathbf{V}_i}^T \theta_V + \sum_{i=1}^N \left((y_i - \mathbf{C}_i^T \mu_{q(\theta)})^2 + \mathbf{C}_i^T \Sigma_{q(\theta)}^T \mathbf{C}_i \right) \exp \left(-\mathbf{C}_{\mathbf{V}_i}^T \theta_V \right) \right. \right. \\ &\quad \left. \left. + \theta_V^T \text{blockdiag} \left(\frac{1}{\sigma_\beta^2} \mathcal{J}_p, \mu_{q(1/\sigma_c^2)} \mathcal{J}_{K_V} \right) \theta_V \right\} \right] = \exp \{ -h(\theta_V | \cdot) \}. \end{aligned} \quad (10)$$

Here $\mu_{q(b)}$ and $\Sigma_{q(b)}$ refer to the elements of the variational mean and covariance of θ corresponding to the random effects vector $\mathbf{b}$. Define

$$B_{q(\sigma_b^2)} = B_b + \parallel \mu_{q(b)} \parallel^2 / 2 + \text{trace} \left(\Sigma_{q(b)} \right) / 2 \text{ and}$$

$$B_{q(\sigma_c^2)} = B_c + \parallel \mu_{q(c)} \parallel^2 / 2 + \text{trace} \left(\Sigma_{q(c)} \right) / 2. \text{ The full derivation of the variational density } q \text{ is in Appendix A.}$$

The lack of a conjugate form for $\theta_V | \cdot$ from (8) carries through to the form of $q^*(\theta_V)$. This makes direct implementation of a variational Bayes approximation difficult since this would involve $(r + K_V + 1)$ -dimensional integration. As in Nott et al. (2012) and Wang and Blei (2013), we consider Laplace approximation to replace $q(\theta_V)$ with a multivariate Gaussian density. Consider the Taylor approximation of the function $h(\theta_V | \cdot)$ from (10) about some value $\mathbf{a}$:

$$\begin{aligned} h(\theta_V | \cdot) &\approx h(\mathbf{a} | \cdot) + (\theta_V - \mathbf{a})^T \mathbf{J}_{\theta_V}(\mathbf{a}) + \frac{1}{2} (\theta_V - \mathbf{a})^T \mathbf{H}_{\theta_V}(\mathbf{a}) (\theta_V - \mathbf{a}), \\ \mathbf{J}_{\theta_V}(\mathbf{a}) &= \frac{\partial}{\partial \mathbf{x}} h(\mathbf{x} | \cdot) |_{\mathbf{x}=\mathbf{a}}, \quad \mathbf{H}_{\theta_V}(\mathbf{a}) = \frac{\partial^2}{\partial^2 \mathbf{x}} h(\mathbf{x} | \cdot) |_{\mathbf{x}=\mathbf{a}}. \end{aligned}$$

Here $\mathbf{J}_{\theta_V}(\mathbf{a})$ corresponds to the Jacobian vector of first derivatives of h with respect to θ_V evaluated at $\mathbf{a}$ and $\mathbf{H}_{\theta_V}(\mathbf{a})$ corresponds to the Hessian matrix containing all partial second derivatives of h evaluated at $\mathbf{a}$. If $\mathbf{a} = \text{argmin } h(\theta_V | \cdot)$, then $\mathbf{J}_{\theta_V}(\mathbf{a}) = \mathbf{0}$ and the approximate form of $q^*(\theta_V)$, which we denote $\tilde{q}^*(\theta_V)$, becomes

$$\tilde{q}^*(\theta_V) \propto \exp \left\{ -\frac{1}{2}(\theta_V - \alpha)^T \mathbf{H}_{\theta_V}(\alpha) (\theta_V - \alpha) \right\}. \quad (11)$$

This implies that θ_V has an approximate variational distribution of the form

$$\theta_V \stackrel{\tilde{q}^*}{\sim} \mathcal{N} \left(\alpha, \left(\mathbf{H}_{\theta_V}(\alpha) \right)^{-1} \right). \quad (12)$$

Alternatives to Laplace approximation, particularly the delta method, could be used to address the non-conjugate problem surrounding θ_V (Wang and Blei, 2013). We choose to use Laplace approximations primarily due to simplicity of implementation, since they require only second- order derivative information whereas the delta method requires third- order.

4.1 Algorithm

Using the q -densities described in (10) and (12), constructing a variational Bayes approximation of our model consists of updating the parameters associated with each q -density until $q(\boldsymbol{\psi})$ has sufficiently approximated $p(\boldsymbol{\psi} | \cdot)$. This is typically done via a coordinate ascent type algorithm (Ormerod and Wand, 2010; Nott et al., 2012; Wang and Blei, 2013). The variational parameters of interest here are denoted as $\boldsymbol{\mu}_{q(\boldsymbol{\theta})}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta})}$, $\boldsymbol{\mu}_{q(\boldsymbol{\theta}_V)}$, $\boldsymbol{\Sigma}_{q(\boldsymbol{\theta}_V)}$, $B_{q(\sigma_b^2)}$, and $B_{q(\sigma_c^2)}$. During each step of the algorithm, the most recent parameters are used and convergence is not assessed until each full cycle has been completed. Given the Inverse Gamma variational distribution of σ_b^2 and σ_c^2 , $\mu_{q(1/\sigma_b^2)} = (A_b + K/2) B_{q(\sigma_b^2)}^{-1}$ and $\mu_{q(1/\sigma_c^2)} = (A_c + K/2) B_{q(\sigma_c^2)}^{-1}$. The terms $\boldsymbol{\mu}_{q(\mathbf{b})}$, $\boldsymbol{\mu}_{q(\mathbf{c})}$, $\boldsymbol{\Sigma}_{q(\mathbf{b})}$, and $\boldsymbol{\Sigma}_{q(\mathbf{c})}$ refer to the portions of the variational means and covariances of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_V$ associated with the random effect vectors.

The calculation of $\mathbf{a}$ at each iteration requires the use of a numerical optimization routine. The speed of this optimization procedure determines the overall speed of the algorithm since the other updates have closed forms. For the results presented here, an implementation of the BFGS Quasi-Newton method found in the built-in `optim` function in R was used for the multivariate optimization problem (Wright and Nocedal, 1999).

The matrices to be inverted in steps 5 and 7 of Algorithm 1 may not be positive definite, perhaps due to poor initial conditions. If an invalid covariance matrix is proposed in a step, we use a ridge adjustment, adding to each diagonal entry of the matrix to be inverted twice the absolute value of its smallest eigenvalue. None of the results presented here required these adjustments.

Algorithm 1 Variational Parameter Updates

```

1: Initialize:  $\mu_{q(\theta)}$ ,  $\Sigma_{q(\theta)}$ ,  $B_{q(\sigma_b^2)}$ , and  $B_{q(\sigma_c^2)}$ 
2: repeat
3:   Compute  $\alpha = \text{argmin}_{\alpha} h(\theta_V | \cdot)$ ,  $\mathbf{H}_{q(\theta_V)}(\alpha)$ 
4:    $\mu_{q(\theta_V)} \leftarrow \alpha$ 
5:    $\Sigma_{q(\theta_V)} \leftarrow (\mathbf{H}_{q(\theta_V)}(\alpha))^{-1}$ 
6:    $B_{q(\sigma_c^2)} \leftarrow B_c + \frac{1}{2} \|\mu_{q(c)}\|^2 + \frac{1}{2} \text{trace}(\Sigma_{q(c)})$ 
7:    $\Sigma_{q(\theta)} \leftarrow \left\{ \text{blockdiag}\left(\frac{1}{\sigma_b^2} \mathbf{I}_p, \mu_{q(1/\sigma_b^2)} \mathbf{I}_K\right) + \mathbf{C}^T \mathbf{I} \mathbf{C} \right\}^{-1}$ 
8:    $\mu_{q(\theta)} \leftarrow \Sigma_{q(\theta)} \mathbf{C}^T \mathbf{I} \mathbf{y}$ 
9:    $B_{q(\sigma_b^2)} \leftarrow B_b + \frac{1}{2} \|\mu_{q(b)}\|^2 + \frac{1}{2} \text{trace}(\Sigma_{q(b)})$ 
10: until Convergence is Reached (See Section 4.2)
11: Construct parameter estimates using means of variational distributions

```

4.2 Convergence Criteria

In fully conjugate variational models such as those in Ormerod and Wand (2010), iteratively updating the variational parameters according to (4) guarantees monotonic increases of $\log p(\mathbf{y}; q)$ as it converges to a (local) maximum. Embedding a Laplace approximation to handle the non-conjugate structure of $q(\theta_V)$ removes this monotonic behavior. Use of an embedded approximation procedure in the context of a variational approximation is generally accepted, however, and can be thought similar to using Metropolis-Hastings within a Gibbs sampler. Nott et al. (2013) formally justify convergence of such an embedded optimization in the context of a fixed form variational Bayes algorithm within a mean field variational Bayes approximation. This is quite similar to our current setting and we therefore conjecture that parameter updates from Algorithm 1 will converge. Empirical results described below support this conjecture.

The evidence lower bound $\log p(\mathbf{y}; q)$ is the objective function monitored to assess convergence. Wang and Blei (2013) suggest monitoring $\|\mu_{q(\theta_V)}\|$ to assess convergence as well. Appendix B contains the derivation of $\log p(\mathbf{y}; q)$.

5 Simulation Example

To illustrate our methodology, consider two simulated datasets of $N = 200$ both with a true mean function of $m(x) = -(1/8)(x - 5)^3 + x$ and with true variance functions of $v_1(x) = \{(1/4)x + (1/2)\}$ or $v_2(x) = \exp \{(1/5)(x - 5)^2\}$, with $x \in [0, 10]$. We fit a model of the form described in (7) using a piecewise quadratic basis ($p = r = 2$). For both the mean and variance levels, $K = K_V = 10$ knots were placed at equally spaced quantiles of x (Ruppert et al., 2003). The hyperparameters associated with the priors for σ_b^2 and σ_c^2 are $A_b = B_b = A_c = B_c = 10^{-5}$. The variance hyperparameters associated with the fixed effects β and δ are $\sigma_\beta^2 = \sigma_\delta^2 = 10^5$.

Before the algorithm is started, an initial fitting of $\mathbf{y}$ under a homoskedastic assumption is performed using the same model structure. The corresponding $\hat{\theta}$ is used as an initial value of $\mu_{q(\theta)}$. The matrix $\Sigma_{q(\theta)}$ is initialized using a diagonal matrix of appropriate scale for the data, and $B_{q(\sigma_b^2)}$ and $B_{q(\sigma_c^2)}$ are initialized to 1/100.

Figure 1 shows the resulting fit of the simulated data under both variance structures using our variational approximation. The approximation closely follows the true mean function in both examples but encounters increased difficulty in areas with higher noise. The shaded areas represent 95% pointwise credible bounds determined by taking a sample of size 10000

from the converged variational distribution of $q^*(\boldsymbol{\theta})$. This measure of uncertainty reflects the non-constant variance structure with wider bands being associated with areas of higher variance. The true mean function $m(x)$ was fully covered by the pointwise credible bounds associated with each variance function.

Figure 2 shows the estimated log variance functions $\log(v_1(x))$ and $\log(v_2(x))$ associated with the data simulated from $m(x)$. The estimated curves associated with our variational approximation (solid lines) closely follow the true log variance functions (dashed lines). The 95% pointwise credible bounds calculated using 10000 draws from the variational distribution $q^*(\boldsymbol{\theta}_V)$ cover the true variance function over the range of x in both cases.

5.1 Comparison to MCMC

We are using an approximate form of the optimal variational density associated with $\boldsymbol{\theta}_V$ and this variational method is itself an approximation. We now consider the quality of our approximation, as compared to the posterior distribution evaluated using a combination of Gibbs sampling and an appropriate multivariate posterior sampling algorithm; in this case, DRAM (see Section 3.1). Figure 3 compares the 95% pointwise credible intervals for both the mean and log variance curves using a simulated sample of 200 points with true curves $m(x)$ and $v_1(x)$. The MCMC procedure was run for 10000 iterations with a burn-in length of 1000. The DRAM step consists of two stages of potential proposal acceptance at each iteration with covariance adaptation every 100 steps. All other model choices (basis functions, knots, etc.) are the same as the previous simulated results.

When comparing the coverage of the true mean curve, both the variational approximation and the MCMC method are very close. This is not surprising given the conjugate structure of the $\boldsymbol{\theta}$ parameter that controls the estimate of $m(x)$. The posterior conditional described in (8) for $\boldsymbol{\theta}$ is multivariate Gaussian as is the variational density $q_{\boldsymbol{\theta}}^*$. The coverage comparison for the log variance curve is more involved. As in Figure 2, the credible intervals from the variational approximation behave smoothly and cover the true function $\log v_1(x)$. The credible bounds from the MCMC procedure have a more pronounced curvature and are narrower in some places. Also, the MCMC bands have some trouble covering $\log v_1(x)$. This behavior can most likely be attributed to the difficulty of the multivariate sampling problem for the posterior conditional distribution $p(\boldsymbol{\theta}_V | \cdot)$. In this case $p(\boldsymbol{\theta}_V | \cdot)$ describes a 13-dimensional distribution from which we have to draw samples. While DRAM performs significantly better than standard Metropolis-Hastings for this problem, it can still have difficulty fully covering the parameter space, especially if the space is complicated and potentially multi-modal. Improvements in fit may be achievable through considerable tuning (modifying adaption length, covariance scaling, etc.) or implementing a more sophisticated algorithm to handle the multivariate sampling problem. These results highlight the additional advantage that variational approximations provide in that they do not require specification of tuning parameters. The coverage comparisons for both the mean and log variance estimates between these two methods suggests that our method does not suffer from the potential underestimation of uncertainty sometimes associated with variational approximations (Rue et al., 2009).

5.2 Computational Performance

Convergence of Algorithm 1 is assessed by monitoring the relative change of $\log p(\mathbf{y}; q)$. Variational algorithms are known to converge very quickly to appropriate estimates (Ormerod and Wand, 2010). Table 1 contains the run times for the variational approximations associated with each variance curve at increasing sample sizes, N . All computations were performed on a MacBook Pro with a 2.3 GHz Intel Core i5 processor and 4 GB of RAM. For all results in Table 1, seven or eight iterations of the algorithm were sufficient to achieve convergence. For comparison, the run time for 10000 iterations of the MCMC procedure described previously are included in Table 1. Our algorithm is slower than those that operate in fully conjugate situations since we have a numerical optimization embedded at each step to estimate α from the most current $h(\theta_V | \cdot)$. However, there is still a dramatic speed increase compared to the MCMC method with run times for the simulated cases being approximately 20 to 60 times faster, depending on sample size, using the variational approximation. This reduction of computational cost coupled with the lack of manual tuning parameters demonstrate the value of these variational approximations. Further computational improvements may be achieved through the use of a custom optimization routine rather than off-the-shelf implementations.

6 Application to Inference from SAXS

The model (7) is directly motivated by the analysis of one-dimensional scattering intensity curves from small angle X-ray scattering (SAXS) experiments (reviewed in Putnam et al. (2007)). The variance of SAXS data depends on angle, due to both characteristics of the molecule and characteristics of the detector. The boundary regions of the detector, associated with larger angles of X-ray scattering, have a smaller signal-to-noise ratio since less scattering occurs at those large angles. This causes greater pixel-to-pixel variation for the intensity values in these regions.

Figure 4 shows data from a single SAXS exposure and the resulting variational approximation fits of the nonparametric mean function, using the approach of Ormerod and Wand (2010) for a model with constant errors (top panel) and using our approach for a model with heteroskedastic errors (bottom panel). The data correspond to the log intensity values of a 7-second exposure of a 4 mg/ml sample of the H2AH2B complex (Isenberg, 1979). The x -axis of Figure 4 represents the radial distance from the center of the sensor. For this problem, we used $K = 40$ equally-spaced knots $\kappa_1, \dots, \kappa_{40}$ over $[0.03, 0.27]$ to evaluate spline functions for both the mean and log variance curves.

The nonparametric regression model uses a truncated quadratic polynomial basis as in (7), with the exception that there is no linear fixed effect term in the mean ($\beta_1 = 0$). This is theoretically justified from physical considerations. Provided the first knot κ_1 is sufficiently close to the origin, a classical analysis yields that expected log intensity at distance x is $\beta_0 - (1/3) (R_g^2) x^2$, where R_g^2 is the squared radius of gyration of the molecule (Guinier, 1939). The use of a truncated quadratic basis is therefore particularly convenient for the estimation of $R_g^2 = -3\beta_2$, where β_2 is the coefficient of the x^2 fixed effect. While truncated polynomial basis functions are highly collinear, we did not experience noticeable

computational problems arising from our choice of this basis. As mentioned in Eilers and Marx (2010), the singular value decomposition of the random effects matrix $\mathbf{Z}$ can aid in understanding the numerical stability of this basis. We do not observe any significant computational or numerical problems in using the truncated quadratic basis for the SAXS data.

The mean estimates for both the constant error and heteroskedastic models fit the data equally well, with mean squared errors of 17.36 and 17.02 respectively. As shown in Figure 4, the uncertainty bands associated with the constant error model are fairly uniform across x . The heteroskedastic version, on the other hand, has narrow uncertainty bounds at the low to mid regions of x , reflecting the pronounced smoothness, while being wider in the large x regions. This behavior better reflects the true uncertainty associated with data about the underlying smooth mean function across the covariate x . This translates to more appropriate uncertainty estimates for the regression coefficients that comprise θ .

Appropriately accounting for the error structure of SAXS data can impact inference on important physical parameters. Here we consider estimation of the radius of gyration, $R_g^2 = -3\beta_2$, and the intercept of the log intensity curve, β_0 , which is known to be linearly related to log molecular weight, given fixed concentration and other experimental parameters (Fischer et al., 2009). Our variational method yields the approximate distributions

7 Discussion

Variational approximations are increasingly popular tools for approximating computationally difficult problems in an efficient manner. Recent work in the literature has focused on the application of mean field variational Bayes methods to non-conjugate problems. This includes the problem presented here: semiparametric regression with heteroskedastic errors. Our work provides a fast variational approximation by leveraging a Laplace approximation to replace the non-conjugate portion of our model with an appropriate Gaussian distribution. Simulation results show that our approximation performs well at fitting both the mean and variance functions, providing posterior inferences comparable to those of a sophisticated MCMC, at much lower computational cost. The ability to fit variance functions accurately is important in accounting for the true uncertainty in nonparametric fits of mean level data, as we have shown in obtaining corrected interval estimates for key physical parameters of H2AH2B using SAXS imaging data.

Under the mixed model framework, extension of our variational approximation only relies on the inclusion of additional random effect variance terms. While not shown explicitly here, our variational approximation readily handles more complex heteroskedastic models. An area of future interest would be further investigation of how to use our variational method for model comparisons in a semiparametric regression context. Nott et al. (2012) highlight the use of the variational lower bound $\log \underline{p}(\mathbf{y}; q)$ for model comparison. We are interested in the use of $\log \underline{p}(\mathbf{y}; q)$ in interpreting the significance of parametric-by-nonparametric interaction terms, which have scientific meaning in SAXS data analysis. For example, a

significant interaction between parametric exposure time and nonparametric angle is suggestive of radiation damage to the molecule in solution.

Radial variance data are typically available alongside radial average data in a SAXS experiment. We are currently extending our variational approximations to handle joint models of radial average data and radial variance data in a complete SAXS experiment, consisting of repeated samples with varying concentrations and exposure times. The goal is to provide a fast, complete framework for the analysis of SAXS data and provide more appropriate uncertainty estimates for inference on structural features.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgment

This research was supported by the Joint NSF/NIGMS Initiative to Support Research in the Area of Mathematical Biology (R01GM096192). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health. This work was completed as a part of Bruce Bugbee's Ph.D. dissertation work while a student at Colorado State University.

A Derivation of q

The following section contains the derivation of the q -densities described in (10).

A.1 q_θ

Using the form of $p(\theta | \cdot)$ from (8), we can derive q_θ

$$\begin{aligned} q_\theta &\propto \exp [E_{-\theta} \{\log p(\theta | \cdot)\}] \\ &\propto \exp \left[E_{-\theta} \left\{ -\frac{1}{2} \left(\theta - \mathbf{M}\mathbf{C}^T \Sigma^{-1} \mathbf{y} \right)^T \mathbf{M}^{-1} \left(\theta - \mathbf{M}\mathbf{C}^T \Sigma^{-1} \mathbf{y} \right) \right\} \right] \\ &\propto \exp \left\{ -\frac{1}{2} E_{-\theta} \left(\theta^T \mathbf{M}^{-1} \theta - \mathbf{y}^T \Sigma^{-1} \mathbf{C} \theta - \theta^T \mathbf{C}^T \Sigma^{-1} \mathbf{y} \right) \right\}. \end{aligned}$$

The quadratic term is

$$\begin{aligned} E_{-\theta} \left(\theta^T \mathbf{M}^{-1} \theta \right) &= E_{-\theta} \left\{ \theta^T \left(\Sigma_\theta^{-1} + \mathbf{C}^T \Sigma^{-1} \mathbf{C} \right) \theta \right\} \\ &= \theta^T \left\{ \text{blockdiag} \left(\frac{1}{\sigma_\beta^2} \mathcal{J}_p, \mu_q (1/\sigma_b^2) \mathcal{J}_K \right) + \mathbf{C}^T \Gamma \mathbf{C} \right\} \theta. \end{aligned}$$

Define $\Sigma_{q(\theta)} = \left\{ \text{blockdiag} \left(\frac{1}{\sigma_\beta^2} \mathcal{J}_p, \mu_q (1/\sigma_b^2) \mathcal{J}_K \right) + \mathbf{C}^T \Gamma \mathbf{C} \right\}^{-1}$. Here $\Gamma = \text{diag}(\gamma_1, \gamma_2, \dots, \gamma_N)$. In this context $\gamma_i = E_{-\theta} \left\{ \exp \left(-\mathbf{C}_{\mathbf{v}_i}^T \theta_{\mathbf{v}_i} \right) \right\}$ corresponds to the moment generating function of $q_{\theta_{\mathbf{v}_i}}$ evaluated at the vector $-\mathbf{C}_{\mathbf{v}_i}^T$.

The expected value of the linear term is

$$E_{-\theta} \left(\mathbf{y}^T \Sigma^{-1} \mathbf{C} \theta \right) = E_{-\theta} \left\{ \sum_{i=1}^N y_i \left(\mathbf{C}_i^T \theta \right) \frac{1}{\sigma_i^2} \right\} = \mathbf{y}^T \Gamma \mathbf{C} \theta.$$

Using these results and completing the square, we arrive at the form of the optimal q_{θ}

$$q_{\theta} \propto \exp \left\{ -\frac{1}{2} \left(\theta - \Sigma_{q(\theta)} \mathbf{C}^T \Gamma \mathbf{y} \right)^T \Sigma_{q(\theta)}^{-1} \left(\theta - \Sigma_{q(\theta)} \mathbf{C}^T \Gamma \mathbf{y} \right) \right\}$$

$$\theta \stackrel{q}{\sim} \mathcal{N} \left(\underbrace{\Sigma_{q(\theta)} \mathbf{C}^T \Gamma \mathbf{y}}_{\mu_{q(\theta)}}, \Sigma_{q(\theta)} \right).$$

A.2 $q_{\sigma_{\mathbf{b}}^2}$

The variational density for $\sigma_{\mathbf{b}}^2$ is

$$q_{\sigma_{\mathbf{b}}^2} \propto \exp \left[E_{-\sigma_{\mathbf{b}}^2} \{ \log p(\sigma_{\mathbf{b}}^2 | \cdot) \} \right]$$

$$\propto \exp \left[E_{-\sigma_{\mathbf{b}}^2} \left\{ - \left(A_{\mathbf{b}} + \frac{K}{2} \right) \log(\sigma_{\mathbf{b}}^2) - \frac{1}{\sigma_{\mathbf{b}}^2} \left(B_{\mathbf{b}} + \|\mathbf{b}\|^2 \right) \right\} \right]$$

$$\propto (\sigma_{\mathbf{b}}^2)^{-A_{\mathbf{b}} - K/2 - 1} \exp \left[-\frac{1}{\sigma_{\mathbf{b}}^2} \left\{ B_{\mathbf{b}} + E_{-\sigma_{\mathbf{b}}^2} (\|\mathbf{b}\|^2) \right\} \right]$$

$$\propto (\sigma_{\mathbf{b}}^2)^{-A_{\mathbf{b}} - K/2 - 1} \exp \left[-\frac{1}{\sigma_{\mathbf{b}}^2} \left\{ B_{\mathbf{b}} + \|\mu_{q(\mathbf{b})}\|^2 + \text{trace}(\Sigma_{q(\mathbf{b})}) \right\} \right]$$

$$\sigma_{\mathbf{b}}^2 \stackrel{q}{\sim} \mathcal{IG} \left(A_{\mathbf{b}} + \frac{K_V}{2}, B_{\mathbf{b}} + \frac{1}{2} \left\{ \underbrace{\|\mu_{q(\mathbf{b})}\|^2 + \text{trace}(\Sigma_{q(\mathbf{b})})}_{B_{q(\sigma_{\mathbf{b}}^2)}} \right\} \right).$$

Here $\mu_{q(\mathbf{b})}$ and $\Sigma_{q(\mathbf{b})}$ refer to the portion of $\mu_{q(\theta)}$ and $\Sigma_{q(\theta)}$ associated with the random effects parameters. Under this variational distribution,

$$E \left[\frac{1}{\sigma_{\mathbf{b}}^2} \right] = \frac{A_{\mathbf{b}} + K/2}{B_{q(\sigma_{\mathbf{b}}^2)}}.$$

A.3 $q_{\sigma_{\mathbf{c}}^2}$

Given the similar structure between $p(\sigma_{\mathbf{b}}^2 | \cdot)$ and $p(\sigma_{\mathbf{c}}^2 | \cdot)$, the derivation of $q_{\sigma_{\mathbf{c}}^2}$ is the same:

$$q_{\sigma_{\mathbf{c}}^2} \propto (\sigma_{\mathbf{c}}^2)^{-A_{\mathbf{c}} - K_V/2 - 1} \exp \left[-\frac{1}{\sigma_{\mathbf{c}}^2} \left\{ B_{\mathbf{c}} + \|\mu_{q(\mathbf{c})}\|^2 + \text{trace}(\Sigma_{q(\mathbf{c})}) \right\} \right]$$

$$\sigma_{\mathbf{c}}^2 \stackrel{q}{\sim} \mathcal{IG} \left(A_{\mathbf{c}} + \frac{K_V}{2}, B_{\mathbf{c}} + \frac{1}{2} \left(\underbrace{\|\mu_{q(\mathbf{c})}\|^2 + \text{trace}(\Sigma_{q(\mathbf{c})})}_{B_{q(\sigma_{\mathbf{c}}^2)}} \right) \right).$$

As before, $\mu_{q(\mathbf{c})}$ and $\Sigma_{q(\mathbf{c})}$ refer to the portion of $\mu_{q(\theta_V)}$ and $\Sigma_{q(\theta_V)}$ associated with the random effects parameters. As with $1/\sigma_{\mathbf{b}}^2$,

$$E \left[\frac{1}{\sigma_c^2} \right] = \frac{A_c + K/2}{B_{q(\sigma_c^2)}}.$$

A.4 q_{θ_V}

As highlighted in (8), the posterior conditional $p(\theta_V | \cdot)$ does not have a known form. Using (4), the optimal q -density for θ_V is

$$\begin{aligned} q_{\theta_V} &\propto \exp \left[E_{-\theta_V} \{ \log p(\theta_V | \cdot) \} \right] \\ &\propto \exp \left(-\frac{1}{2} \left[\sum_{i=1}^N \mathbf{C}_{\mathbf{v}_i}^T \theta_V + \sum_{i=1}^N \left\{ \left(y_i - \mathbf{C}_i^T \mu_{q(\theta)} \right)^2 + \mathbf{C}_i^T \Sigma_{q(\theta)} \mathbf{C}_i \right\} \exp \left(-\mathbf{C}_{\mathbf{v}_i}^T \theta_V \right) \right. \right. \\ &\quad \left. \left. + \theta_V^T \text{blockdiag} \left(\frac{1}{\sigma_\delta^2} \mathcal{J}_r, \mu_{q(1/\sigma_c^2)} \mathcal{J}_{K_V} \right) \theta_V \right] \right). \end{aligned}$$

B Derivation of $\log p(\mathbf{y}; \mathbf{q})$

Let $\Gamma(x)$ denote the standard gamma function. Recall that the density $q(\boldsymbol{\psi})$ is assumed to have the product density structure described in Section 4. Then

$$\begin{aligned} \log p(\mathbf{y}; \mathbf{q}) &= \int_{\boldsymbol{\psi}} q(\boldsymbol{\psi}) [\log p(\mathbf{y}, \boldsymbol{\psi}) - \log q(\boldsymbol{\psi})] d\boldsymbol{\psi} \\ &= \int_{\boldsymbol{\psi}} q(\boldsymbol{\psi}) [\log p(\mathbf{y} | \boldsymbol{\psi}) + \log p(\boldsymbol{\psi}) - \log q(\boldsymbol{\psi})] d\boldsymbol{\psi} \\ &= \int_{\boldsymbol{\psi}} q(\boldsymbol{\psi}) [\log p(\mathbf{y} | \boldsymbol{\psi}) + \log p(\theta) + \log p(\sigma_b^2) + \log p(\theta_V) + \log p(\sigma_c^2) - \log q(\boldsymbol{\psi})] d\boldsymbol{\psi} \\ &= -\frac{1}{2} \log(2\pi) N - \frac{r}{2} \log(\sigma_\beta^2) - \frac{r}{2} \log(\sigma_\delta^2) + A_b \log(B_b) - \log(\Gamma(A_b)) \\ &\quad + A_c \log(B_c) - \log(\Gamma(A_c)) + \frac{1}{2} \log(|\Sigma_{q(\theta)}|) + \frac{1}{2} \log(|\Sigma_{q(\theta_V)}|) \\ &\quad - \left(A_b + \frac{K}{2} \right) \log(B_q(\sigma_b^2)) - \left(A_c + \frac{K_V}{2} \right) \log(B_{q(\sigma_c^2)}) \\ &\quad + \log\left(\Gamma\left(A_c + \frac{K_V}{2}\right)\right) + \log\left(\Gamma\left(A_b + \frac{K}{2}\right)\right) \\ &\quad + \frac{1}{2\sigma_\beta^2} (\|\mu_{q(\beta)}\|^2 + \text{trace}(\Sigma_{q(\beta)})) + \frac{1}{2\sigma_\delta^2} (\|\mu_{q(\delta)}\|^2 + \text{trace}(\Sigma_{q(\delta)})) \\ &\quad - \frac{1}{2} \sum_{i=1}^N \left\{ \left(y_i - \mathbf{C}_i^T \mu_{q(\theta)} \right)^2 + \mathbf{C}_i^T \Sigma_{q(\theta)} \mathbf{C}_i \right\} \exp \left(-\mathbf{C}_{\mathbf{v}_i}^T \mu_{q(\theta_V)} + \frac{1}{2} \mathbf{C}_{\mathbf{v}_i}^T \Sigma_{q(\theta)} \mathbf{C}_{\mathbf{v}_i} \right) \\ &\quad + \frac{p+K}{2} + \frac{r+K_V}{2}. \end{aligned}$$

References

- Attias H. A Variational Bayesian Framework for Graphical Models. *Advances in Neural Information Processing Systems*. 2000; 12:209–215.
- Bishop, CM. *Pattern Recognition and Machine Learning*. Information Science and Statistics, Springer; 2006.
- Blei DM, Lafferty JD. A Correlated Topic Model of Science. *The Annals of Applied Statistics*. 2007; 1:17–35.
- Breidt FJ, Erciulescu A, van der Woerd M. Autocovariance Structures for Radial Averages in Small-Angle X-Ray Scattering Experiments. *Journal of Time Series Analysis*. 2012; 33:704–717. [PubMed: 23355752]
- Carlin BP, Polson NG. Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler. *Canadian Journal of Statistics*. 1991; 19:399–405.

- Casella G, George EI. Explaining the Gibbs Sampler. *The American Statistician*. 1992; 46:167–174.
- Chan D, Kohn R, Nott D, Kirby C. Locally Adaptive Semiparametric Estimation of the Mean and Variance Functions in Regression Models. *Journal of Computational and Graphical Statistics*. 2006; 15:915–936.
- Crainiceanu CM, Ruppert D, Carroll RJ, Joshi A, Goodner B. Spatially Adaptive Bayesian Penalized Splines with Heteroscedastic Errors. *Journal of Computational and Graphical Statistics*. 2007; 16:265–288.
- Eilers PH, Marx BD. Splines, Knots, and Penalties. *Wiley Interdisciplinary Reviews: Computational Statistics*. 2010; 2:637–653.
- Eilers PHC, Marx BD. Flexible Smoothing with B-splines and Penalties. *Statistical Science*. 1996; 11:89–102.
- Faes C, Ormerod J, Wand M. Variational Bayesian Inference for Parametric and Nonparametric Regression with Missing Data. *Journal of the American Statistical Association*. 2011; 106:959–971.
- Fischer H, Oliveira Neto M. d. Napolitano H, Polikarpov I, Craievich A. Determination of the Molecular Weight of Proteins in Solution from a Single Small-Angle X-Ray Scattering Measurement on a Relative Scale. *Journal of Applied Crystallography*. 2009; 43:101–109.
- Guinier A. La Diffraction des Rayons X aux Très Petits Angles: Application à l'Étude de Phénomènes Ultramicroscopiques. *Annales de Physique*. 1939; 12:161–237.
- Haario H, Laine M, Mira A, Saksman E. DRAM: Efficient Adaptive MCMC. *Statistics and Computing*. 2006; 16:339–354.
- Isenberg I. Histones. *Annual Review of Biochemistry*. 1979; 48:159–191.
- Jaakkola, T.; Jordan, M. A Variational Approach to Bayesian Logistic Regression Models and Their Extensions. *Sixth International Workshop on Artificial Intelligence and Statistics*; 1997.
- Knowles DA, Minka T. Non-Conjugate Variational Message Passing for Multinomial and Binary Regression. *Advances in Neural Information Processing Systems*. 2011:1701–1709.
- Kullback S, Leibler RA. On Information and Sufficiency. *The Annals of Mathematical Statistics*. 1951; 22:79–86.
- Minka, T.; Winn, J.; Guiver, J.; Knowles, D. *Infer.NET 2.4*. Microsoft Research Cambridge. 2010. <http://research.microsoft.com/infernet>
- Nott DJ, Tran M-N, Kuk AY, Kohn R. Efficient Variational Inference for Generalized Linear Mixed Models with Large Datasets. *arXiv preprint arXiv:1307.7963*. 2013
- Nott DJ, Tran M-N, Leng C. Variational Approximation for Heteroscedastic Linear Models and Matching Pursuit Algorithms. *Statistics and Computing*. 2012; 22:497–512.
- Ormerod J, Wand M. Explaining Variational Approximations. *The American Statistician*. 2010; 64:140–153.
- Pham TH, Ormerod JT, Wand M. Mean Field Variational Bayesian Inference for Nonparametric Regression with Measurement Error. *Computational Statistics & Data Analysis*. 2013; 68:375–387.
- Putnam C, Hammel M, Hura G, Tainer J. X-Ray Solution Scattering (SAXS) Combined with Crystallography and Computation: Defining Accurate Macromolecular Structures, Conformations and Assemblies in Solution. *Quarterly Reviews of Biophysics*. 2007; 40:191–285. [PubMed: 18078545]
- Rue H, Martino S, Chopin N. Approximate Bayesian Inference for Latent Gaussian models by Using Integrated Nested Laplace Approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2009; 71:319–392.
- Ruppert, D.; Wand, MP.; Carroll, RJ. *Semiparametric Regression*. Vol. 12. Cambridge University Press; 2003.
- Salimans T, Knowles DA. Fixed-Form Variational Posterior Approximation Through Stochastic Linear Regression. *Bayesian Analysis*. 2013; 8:837–882.
- Tan LS, Nott DJ. Variational Inference for Generalized Linear Mixed Models Using Partially Noncentered Parametrizations. *Statistical Science*. 2013; 28:168–188.

- Titterton D. Bayesian Methods for Neural Networks and Related Models. *Statistical Science*. 2004; 19:128–139.
- Wainwright MJ, Jordan MI. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*. 2008; 1:1–305.
- Wand M, Ormerod J. On Semiparametric Regression with O’Sullivan Penalized Splines. *Australian & New Zealand Journal of Statistics*. 2008; 50:179–198.
- Wand MP, Ormerod JT, Padoan SA, Frühwirth R. Mean Field Variational Bayes for Elaborate Distributions. *Bayesian Analysis*. 2011; 6:1–48. [PubMed: 22247752]
- Wang C, Blei DM. Variational Inference in Nonconjugate Models. *The Journal of Machine Learning Research*. 2013; 14:1005–1031.
- Wang S, Wand M. Using Infer.NET for Statistical Analyses. *The American Statistician*. 2011; 65:115.
- Wright, S.; Nocedal, J. *Numerical Optimization*. Vol. 2. Springer; New York: 1999.

Figure 1.

$N=200$ simulated observations with true mean function $m(x) = -(1/8)(x-5)^3 + x$ (dashed line in each plot). Variance functions are $v_1(x) = \{(1/4)x + (1/2)\}^3$ (top) and $v_2(x) = \exp\{(1/5)(x-5)^2\}$ (bottom). Solid lines represent the approximate fits of $m(x)$ using the variational method described in Algorithm 1. The shaded area corresponds to the 95% pointwise credible bounds calculated from 10000 draws from the variational distribution $q^*(\theta)$.

Figure 2.

The solid lines represent the estimated log variance functions associated with the variational approximation detailed in Figure 1. The shaded area corresponds to the 95% pointwise credible bounds calculated from 10000 draws from the variational distribution $q^*(\theta_V)$. The true log variance functions are represented by dashed lines.

Figure 3.

The solid light gray regions represent the pointwise 95% credible bounds drawn from the approximating variational distributions q_{θ} and $\tilde{q}_{\theta_V}$ for a simulated data set of 200 observations with $m(x)$ and $v_1(x)$ as the true mean and variance functions. The line shaded regions correspond to the resulting pointwise posterior 95% credible bounds taken from the hybrid Gibbs-DRAM MCMC procedure using the posterior conditionals described in (8) and run for 10000 iterations with a 10% burn-in length.

Figure 4.

Constant error (top) and heteroskedastic (bottom) variational approximation fits of experimental SAXS data from a single 7-second exposure of a 4 mg/ml sample of the H2AH2B complex. The gray bands represent 95% pointwise credible intervals generated from 10000 draws from the converged approximate variational distributions. The knots $\kappa_1, \dots, \kappa_{40}$ are represented by small tick marks along the x -axis.

Table 1

Run times (seconds) for the variational approximations and MCMC procedures described in Section 5 at various sample sizes. The MCMC procedure was run for 10000 iterations in each case. All variational approximations required seven or eight iterations to achieve convergence.

Method	Sample Size (N)			
	200	400	800	1600
Variational approximation for $v_1(x)$	10.68	20.02	39.35	98.67
Variational approximation for $v_2(x)$	10.73	21.18	42.18	95.30
MCMC for $v_1(x)$	594.75	657.99	872.55	1702.41
MCMC for $v_2(x)$	592.98	660.51	874.53	1897.50

Table 2

Estimated intercept $\hat{\beta}_0$ and squared radius of gyration $\widehat{R}_g^2$ of the log intensity curve for the complex H2AH2B arising from variational approximations of both constant-error and heteroskedastic models. The corresponding credible bounds are derived using the quantiles of the Gaussian distributions described in §6.

		Constant Errors	Heteroskedastic
$\hat{\beta}_0$	Estimate	6.16	6.17
	95% Credible Bounds	[5.95, 6.37]	[6.15, 6.18]
	Interval Width	0.42	0.03
$\widehat{R}_g^2$	Estimate	553.48	535.40
	95% Credible Bounds	[205.66, 897.39]	[484.08, 586.74]
	Interval Width	691.73	102.66