Journal of Machine Learning Research 15 (2014) 1351-1369 Submitted 7/13; Revised 11/13; Published 4/14

Fully Simplified Multivariate Normal Updates in Non-Conjugate Variational Message Passing

Matt P. Wand MATT.WAND@UTS.EDU.AU
School of Mathematical Sciences
University of Technology, Sydney
P.O. Box 123, Broadway NSW 2007, Australia

Editor: David M. Blei

Abstract

Fully simplified expressions for Multivariate Normal updates in non-conjugate variational message passing approximate inference schemes are obtained. The simplicity of these expressions means that the updates can be achieved very efficiently. Since the Multivariate Normal family is the most common for approximating the joint posterior density function of a continuous parameter vector, these fully simplified updates are of great practical benefit.

Keywords: Bayesian computing, graphical models, matrix differential calculus, mean field variational Bayes, variational approximation

1. Introduction

Recently Knowles and Minka (2011) proposed a prescription for handling non-conjugate exponential family factors in variational message passing approximate inference schemes. Dubbed *non-conjugate variational message passing*, it widens the scope of tractable models for variational message passing and mean field variational Bayes in general. For a given exponential family factor, the non-conjugate variational message passing updates depend on the inverse covariance matrix of the natural statistic and derivatives of the non-entropy component of the Kullback-Leibler divergence.

The Multivariate Normal distribution is the most common multivariate exponential family distribution and a prime candidate for approximating the joint posterior density function of a continuous parameter vector, such as a set of regression coefficients. Knowles and Minka (2011) provide formulae for Univariate Normal updates, which correspond to less accurate diagonal covariance matrix approximations to joint posterior density functions. However, when combined with the derived variable infrastructure described Appendix A of Minka and Winn (2008), the Univariate Normal updates in Knowles and Minka (2011) are able to produce full covariance matrix Multivariate Normal approximations for regression models. This fact is utilized by the Infer.NET computational framework (Minka et al., 2013), although the mathematical description of the updates is somewhat verbose. This aspect hinders extension to more complicated models, including those not supported by Infer.NET.

Recently, Tan and Nott (2013) utilized non-conjugate variational message passing for approximate Bayesian inference in hierarchical generalized linear mixed models. Their numerical studies showed that their variational algorithms can achieve high levels of accuracy.

This accuracy is partly due to their use of Multivariate Normal, rather than Univariate Normal, factors.

This article's main contribution is full simplification of the inverse covariance matrix of the natural statistic and then to show that the updates admit a particularly simple form in terms of derivatives with respect to the common Multivariate Normal parameters, that is, the mean and covariance matrix. When combined with an additional novel matrix result, this article's second theorem, non-conjugate mean field variational Bayes algorithms involving Multivariate Normal updates are straightforward to derive and implement. This explicitness allows much easier accommodation of Multivariate Normal posterior density functions within the non-conjugate variational message passing framework. Algorithm 3 of Tan and Nott (2013) relies on Theorems 1 and 2, presented in Section 4. This leads to considerable computational efficiency for the methodology in Tan and Nott (2013).

Non-conjugate variational message passing (Knowles and Minka, 2011) is one of several recent contributions aimed at widening the set of models that can be handled via the mean field variational Bayes paradigm. Others include Braun and McAuliffe (2010), Wand et al. (2011) and Wang and Blei (2013).

Section 2 lays out notation needed for the main theorems, which are presented in Section 4. The utility of these theorems is then illustrated in Section 5 for a Bayesian Poisson mixed model and a heteroscedastic additive model. A series of appendices contains proofs of the theorems and other mathematical details.

2. Notation

The main results makes ample use of the matrix differential calculus technology of Magnus and Neudecker (1999). Therefore, I mainly adhere to their notation.

2.1 The vec, vech and Duplication Matrix Notations

If $\boldsymbol{A}$ is a $d \times d$ matrix then $\text{vec}(\boldsymbol{A})$ denotes the $d^2 \times 1$ vector obtained by stacking the columns of $\boldsymbol{A}$ underneath each other in order from left to right. Also, $\text{vech}(\boldsymbol{A})$ denotes the $\frac{1}{2}\,d(d+1) \times 1$ vector obtained from $\text{vec}(\boldsymbol{A})$ by eliminating each of the above-diagonal entries of $\boldsymbol{A}$. For example,

$$\text{vec}\left(\begin{bmatrix} 5 & 2 \\ 9 & 4 \end{bmatrix}\right) = \begin{bmatrix} 5 \\ 9 \\ 2 \\ 4 \end{bmatrix} \quad \text{while} \quad \text{vech}\left(\begin{bmatrix} 5 & 2 \\ 9 & 4 \end{bmatrix}\right) = \begin{bmatrix} 5 \\ 9 \\ 4 \end{bmatrix}.$$

If $\boldsymbol{A}$ is a symmetric, but otherwise arbitrary $d \times d$ matrix, then $\text{vech}(\boldsymbol{A})$ contains each of the distinct entries of $\boldsymbol{A}$ whereas $\text{vec}(\boldsymbol{A})$ repeats the off-diagonal entries. It follows that there is a unique $d^2 \times \frac{1}{2}d(d+1)$ matrix $\boldsymbol{D}_d$ of zeros and ones such that

$$\boldsymbol{D}_d\,\text{vech}(\boldsymbol{A}) = \text{vec}(\boldsymbol{A}) \quad \text{for} \quad \boldsymbol{A} = \boldsymbol{A}^T$$

and is called the *duplication matrix of order d*. The Moore-Penrose inverse of $\boldsymbol{D}_d$ is

$$\boldsymbol{D}_d^+ \equiv (\boldsymbol{D}_d^T\,\boldsymbol{D}_d)^{-1}\boldsymbol{D}_d^T.$$

Note that

$$\boldsymbol{D}_d^+ \operatorname{vec}(\boldsymbol{A}) = \operatorname{vech}(\boldsymbol{A}) \quad \text{for} \quad \boldsymbol{A} = \boldsymbol{A}^T.$$

The simplest non-trivial examples of $\boldsymbol{D}_d$ and $\boldsymbol{D}_d^+$ are

$$\boldsymbol{D}_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{D}_2^+ = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Note that, for general d, $\boldsymbol{D}_d$ can be obtained via the `duplication.matrix()` function in the package `matrixcalc` (Novomestky, 2008) within the R computing environment (R Development Core Team, 2013).

If $\boldsymbol{a}$ is a $d^2 \times 1$ vector then $\operatorname{vec}^{-1}(\boldsymbol{a})$ is defined to be the $d \times d$ matrix formed from listing the entries of $\boldsymbol{a}$ in a column-wise fashion in order from left to right. Note that vec^{-1} is the usual function inverse when the domain of vec is restricted to square matrices. In particular,

$$\operatorname{vec}^{-1}(\operatorname{vec}(\boldsymbol{A})) = \boldsymbol{A} \quad \text{for } d \times d \text{ matrices } \boldsymbol{A}$$

and

$$\operatorname{vec}(\operatorname{vec}^{-1}(\boldsymbol{a})) = \boldsymbol{a} \quad \text{for } d^2 \times 1 \text{ vectors } \boldsymbol{a}.$$

There are numerous identities involving vec, vech, $\boldsymbol{D}_d$ and $\boldsymbol{D}_d^+$, and some of these are given in Chapter 3 of Magnus and Neudecker (1999). One that is relevant to the current article is:

Lemma 1 *If $\boldsymbol{A}$ is a symmetric $d \times d$ matrix then*

$$\operatorname{vec}(\boldsymbol{A}) = \boldsymbol{D}_d^{+T} \boldsymbol{D}_d^T \operatorname{vec}(\boldsymbol{A}).$$

2.2 The diagonal and diag Notations

If $\boldsymbol{A}$ is a $d \times d$ matrix then diagonal($\boldsymbol{A}$) denotes the $d \times 1$ vector containing the diagonal entries of $\boldsymbol{A}$. If $\boldsymbol{a}$ is a $d \times 1$ vector then diag($\boldsymbol{a}$) is the $d \times d$ matrix with the entries of $\boldsymbol{a}$ on the diagonal and all other entries equal to zero. For example,

$$\operatorname{diagonal}\left(\begin{bmatrix} 8 & 1 & -7 \\ 3 & 6 & 24 \\ -4 & 11 & -9 \end{bmatrix}\right) = \begin{bmatrix} 8 \\ 6 \\ -9 \end{bmatrix} \quad \text{and} \quad \operatorname{diag}\left(\begin{bmatrix} -4 \\ 7 \\ 31 \end{bmatrix}\right) = \begin{bmatrix} -4 & 0 & 0 \\ 0 & 7 & 0 \\ 0 & 0 & 31 \end{bmatrix}.$$

2.3 Derivative Vector and Hessian Matrix Notation

Let f be a $\mathbb{R}^p$-valued function with argument $\boldsymbol{x} \in \mathbb{R}^d$. The *derivative vector* of f with respect to $\boldsymbol{x}$, $\mathsf{D}_{\boldsymbol{x}} f$, is the $p \times d$ matrix whose (i, j) entry is

$$\frac{\partial f_i(\boldsymbol{x})}{\partial x_j}.$$

where $f_i(\boldsymbol{x})$ is the ith entry of $f(\boldsymbol{x})$ and x_j is the jth entry of $\boldsymbol{x}$. For example

$$\begin{aligned}\mathsf{D}_{\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]} \begin{bmatrix} \tan(x_1 + 7x_2) \\ 3x_1^4(8 + 9x_2^3) \end{bmatrix} &= \begin{bmatrix} \frac{\partial}{\partial x_1}\{\tan(x_1 + 7x_2)\} & \frac{\partial}{\partial x_2}\{\tan(x_1 + 7x_2)\} \\ \frac{\partial}{\partial x_1}\{3x_1^4(8 + 9x_2^3)\} & \frac{\partial}{\partial x_2}\{3x_1^4(8 + 9x_2^3)\} \end{bmatrix} \\ &= \begin{bmatrix} \sec^2(x_1 + 7x_2) & 7\sec^2(x_1 + 7x_2) \\ 12x_1^3(8 + 9x_2^3) & 81x_1^4x_2^2 \end{bmatrix}.\end{aligned}$$

In the case $p = 1$, the *Hessian matrix* of f with respect to $\boldsymbol{x}$, $\mathsf{H}_{\boldsymbol{x}} f$, is the $d \times d$ matrix

$$\mathsf{H}_{\boldsymbol{x}} f = \mathsf{D}_{\boldsymbol{x}}\{(\mathsf{D}_{\boldsymbol{x}} f)^T\}.$$

3. Non-Conjugate Variational Message Passing

Non-conjugate variational message passing (Knowles and Minka, 2011) is an extension of mean field variational Bayes (e.g. Wainwright and Jordan, 2008) where, due to difficulties arising from non-conjugacy, one or more density functions is forced to have a particular exponential family distribution.

Consider a hierarchical Bayesian model with data vector $\boldsymbol{y}$ and parameter vectors $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$. Mean field variational Bayes approximates the joint posterior density function $p(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{y})$ by

$$q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1) \cdots q_{\boldsymbol{\theta}_M}(\boldsymbol{\theta}_M)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}), \tag{1}$$

where $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_M\}$ is a partition of $\boldsymbol{\theta}$ and each subscripted q is an unrestricted density function. The solutions satisfy

$$\begin{aligned} &q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \propto \exp[E_{q(-\boldsymbol{\theta}_i)}\{\log p\,(\boldsymbol{\theta}_i|\boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i, \boldsymbol{\phi})\}], \quad 1 \le i \le M, \\ &q^*_{\boldsymbol{\phi}}(\boldsymbol{\phi}) \propto \exp[E_{q(-\boldsymbol{\phi})}\{\log p\,(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\theta})\}], \end{aligned}$$

where $\boldsymbol{\theta}\backslash\boldsymbol{\theta}_i$ means $\boldsymbol{\theta}$ with $\boldsymbol{\theta}_i$ excluded and $E_{q(-\boldsymbol{\theta}_i)}$ denotes expectation with respect to the q-densities of all parameters except $\boldsymbol{\theta}_i$. A similar definition applies to $E_{q(-\boldsymbol{\phi})}$.

In the event that $E_{q(-\boldsymbol{\phi})}\{\log p\,(\boldsymbol{\phi}|\boldsymbol{y}, \boldsymbol{\theta})\}$ is intractable, non-conjugate variational message passing offers a way out by replacing (1) with

$$q_{\boldsymbol{\theta}_1}(\boldsymbol{\theta}_1) \cdots q_{\boldsymbol{\theta}_M}(\boldsymbol{\theta}_M)\, q_{\boldsymbol{\phi}}(\boldsymbol{\phi}; \boldsymbol{\eta}),$$

where $q_{\boldsymbol{\phi}}(\boldsymbol{\phi}; \boldsymbol{\eta})$ is an exponential family density function with *natural parameter vector* $\boldsymbol{\eta}$ and *natural statistic* $\boldsymbol{T}(\boldsymbol{\phi})$. Then, with backing from Theorem 1 of Knowles and Minka (2011), the optimal densities $q^*(\theta_1), \ldots, q^*(\theta_M)$ and $q^*(\boldsymbol{\phi}; \boldsymbol{\eta})$ may be found using

$$\begin{aligned} &q^*_{\boldsymbol{\theta}_i}(\boldsymbol{\theta}_i) \propto \exp[E_{q(-\boldsymbol{\theta}_i)}\{\log p(\theta_i|\boldsymbol{y}, \boldsymbol{\theta}\backslash\boldsymbol{\theta}_i, \boldsymbol{\phi})\}], \quad 1 \le i \le M, \\ &\boldsymbol{\eta} \leftarrow [\,\mathrm{var}\{\boldsymbol{T}(\boldsymbol{\phi})\}]^{-1}\, [\mathsf{D}_{\boldsymbol{\eta}}\, E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\{\log p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}]^T. \end{aligned} \tag{2}$$

Here and elsewhere $\mathrm{var}(\boldsymbol{v})$ denotes the covariance matrix of a random vector $\boldsymbol{v}$. As pointed out in Knowles and Minka (2011), the graphical structure of the hierarchical Bayesian model can be used to provide a simpler expression for $\mathsf{D}_{\boldsymbol{\eta}}\, E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\{\log p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})\}$ that only depends on factors of $p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi})$ involving $\boldsymbol{\phi}$.

3.1 Multivariate Normal Factor

Now consider the special case where $q(\boldsymbol{\phi};\boldsymbol{\eta})$ corresponds to a d-dimensional Multivariate Normal density function. Then the natural statistic (defined in Section 4) is

$$\boldsymbol{T}(\boldsymbol{\phi}) \equiv \left[\begin{array}{c} \boldsymbol{\phi} \\ \mathrm{vech}(\boldsymbol{\phi}\,\boldsymbol{\phi}^T) \end{array} \right].$$

Since $\boldsymbol{T}(\boldsymbol{\phi})$ has $d + d(d+1)/2$ entries, the number of entries in $\mathrm{var}\{\boldsymbol{T}(\boldsymbol{\phi})\}$ is quartic in d. Consequently, for large d, the $\boldsymbol{\eta}$ update in (2) is numerically challenging if done directly. In Section 4 I present theoretical results that allow explicit updating without the need for inversion of $\mathrm{var}\{\boldsymbol{T}(\boldsymbol{\phi})\}$. I also present results in terms of the common Multivariate Normal parametrization, involving mean vectors and covariance matrices.

4. Main Results

Consider a generic Multivariate Normal $d \times 1$ random vector $\boldsymbol{x}$

$$\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{3}$$

Then the density function of $\boldsymbol{x}$ is

$$\begin{aligned} p(\boldsymbol{x}) &= (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\} \\ &= \exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta} - A(\boldsymbol{\eta}) - \tfrac{d}{2}\log(2\pi)\}. \end{aligned}$$

Here

$$\boldsymbol{T}(\boldsymbol{x}) \equiv \left[\begin{array}{c} \boldsymbol{x} \\ \mathrm{vech}(\boldsymbol{x}\,\boldsymbol{x}^T) \end{array} \right], \quad \boldsymbol{\eta} \equiv \left[\begin{array}{c} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{array} \right] \equiv \left[\begin{array}{c} \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\ -\frac{1}{2}\,\boldsymbol{D}_d^T\mathrm{vec}(\boldsymbol{\Sigma}^{-1}) \end{array} \right] \tag{4}$$

defines the *natural statistic* and *natural parameter* pairing and

$$A(\boldsymbol{\eta}) = -\tfrac{1}{4}\,\boldsymbol{\eta}_1^T\left\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1}\boldsymbol{\eta}_1 - \tfrac{1}{2}\log\left|-2\,\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right|$$

is the *log-partition* function.

Note that the inverse of the natural parameter transformation is

$$\left\{ \begin{array}{ll} \boldsymbol{\mu} = & -\frac{1}{2}\left\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1}\boldsymbol{\eta}_1 \\ \boldsymbol{\Sigma} = & -\frac{1}{2}\left\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\right\}^{-1} \end{array} \right. \tag{5}$$

and can be derived from (4) using Lemma 1.

Theorem 1 *Consider the $d \times 1$ random vector $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with natural statistic vector $T(\boldsymbol{x})$ and natural parameter vector $\boldsymbol{\eta}$ given by (4) and define*

$$\boldsymbol{U} \equiv \left\{\mathsf{D}_{\boldsymbol{\eta}}\left[\begin{array}{c} \boldsymbol{\mu} \\ \mathrm{vec}(\boldsymbol{\Sigma}) \end{array} \right]\right\}^T, \quad \boldsymbol{V} \equiv \mathrm{var}\{T(\boldsymbol{x})\} = \mathsf{H}_{\boldsymbol{\eta}}A(\boldsymbol{\eta}),$$

$$\boldsymbol{M} \equiv 2\,\boldsymbol{D}_d^+(\boldsymbol{\mu} \otimes \boldsymbol{I}_d) \quad \textit{and} \quad \boldsymbol{S} \equiv 2\boldsymbol{D}_d^+(\boldsymbol{\Sigma} \otimes \boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}.$$

Then

$$(a) \qquad \boldsymbol{U} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{M}\boldsymbol{\Sigma} & \boldsymbol{S}\boldsymbol{D}_d^T \end{bmatrix},$$

$$(b) \qquad \boldsymbol{S}^{-1} = \tfrac{1}{2}\boldsymbol{D}_d^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\boldsymbol{D}_d,$$

$$(c) \qquad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{M}^T \\ \boldsymbol{M}\boldsymbol{\Sigma} & \boldsymbol{S} + \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T \end{bmatrix},$$

$$(d) \qquad \boldsymbol{V}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} + \boldsymbol{M}^T\boldsymbol{S}^{-1}\boldsymbol{M} & -\boldsymbol{M}^T\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M} & \boldsymbol{S}^{-1} \end{bmatrix},$$

$$(e) \qquad \boldsymbol{V}^{-1}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{M}^T\boldsymbol{D}_d^T \\ \boldsymbol{0} & \boldsymbol{D}_d^T \end{bmatrix}$$

and

$$(f) \qquad \boldsymbol{V}^{-1}\boldsymbol{U}\begin{bmatrix} \boldsymbol{g} \\ \mathrm{vec}(\boldsymbol{G}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{g} - 2\,\boldsymbol{G}\boldsymbol{\mu} \\ \boldsymbol{D}_d^T\mathrm{vec}(\boldsymbol{G}) \end{bmatrix}$$

for every $d \times 1$ vector $\boldsymbol{g}$ and symmetric $d \times d$ matrix $\boldsymbol{G}$.

Appendix A contains a proof of Theorem 1.

Let s be a smooth function of $\boldsymbol{\eta}$, the natural parameter vector in a Multivariate Normal factor, and consider an iterative scheme with updates of the form

$$\boldsymbol{\eta} \leftarrow \boldsymbol{V}^{-1}(\mathsf{D}_{\boldsymbol{\eta}}\, s)^T. \tag{6}$$

Note that the update in (2) is a special case of (6) with $s(\boldsymbol{\eta}) = \mathsf{D}_{\boldsymbol{\eta}}\, E_{q(\boldsymbol{\theta},\boldsymbol{\phi})}\{\log\, p(\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\eta})\}$. By the chain rule of matrix differential calculus (Theorem 8, Chapter 5, of Magnus and Neudecker, 1999)

$$\boldsymbol{V}^{-1}(\mathsf{D}_{\boldsymbol{\eta}}\, s)^T = \boldsymbol{V}^{-1}\left[\left\{\mathsf{D}_{\left[\begin{smallmatrix}\boldsymbol{\mu}\\ \mathrm{vec}(\boldsymbol{\Sigma})\end{smallmatrix}\right]}\, s\right\}\left\{\mathsf{D}_{\boldsymbol{\eta}}\begin{bmatrix}\boldsymbol{\mu}\\ \mathrm{vec}(\boldsymbol{\Sigma})\end{bmatrix}\right\}\right]^T = \boldsymbol{V}^{-1}\boldsymbol{U}\begin{bmatrix}(\mathsf{D}_{\boldsymbol{\mu}}\, s)^T \\ (\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\, s)^T\end{bmatrix}.$$

Let $(\boldsymbol{\mu}_{\mathrm{old}}, \boldsymbol{\Sigma}_{\mathrm{old}})$ and $(\boldsymbol{\mu}_{\mathrm{new}}, \boldsymbol{\Sigma}_{\mathrm{new}})$, respectively, denote the old and new values of $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in the updating scheme (6). Then, it follows from Theorem 1(f) that

$$\begin{aligned} \boldsymbol{\Sigma}_{\mathrm{new}}^{-1}\boldsymbol{\mu}_{\mathrm{new}} &= \left[(\mathsf{D}_{\boldsymbol{\mu}}\, s)^T - 2\,\mathrm{vec}^{-1}\left((\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\, s)^T\right)\boldsymbol{\mu}\right]_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\mathrm{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\mathrm{old}}} \\ \text{and} \quad \boldsymbol{D}_d^T\mathrm{vec}(-\tfrac{1}{2}\boldsymbol{\Sigma}_{\mathrm{new}}^{-1}) &= \left[\boldsymbol{D}_d^T(\mathsf{D}_{\mathrm{vec}(\boldsymbol{\Sigma})}\, s)^T\right]_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\mathrm{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\mathrm{old}}}. \end{aligned}$$

The mean and covariance parameter updates are therefore given by

$$\begin{aligned} \boldsymbol{\Sigma}_{\text{new}} &= \{-2\,\text{vec}^{-1}(\{[\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\, s]_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}}\}^T)\}^{-1} \\ \text{and} \quad \boldsymbol{\mu}_{\text{new}} &= \boldsymbol{\mu}_{\text{old}} + \boldsymbol{\Sigma}_{\text{new}}([\mathsf{D}_{\boldsymbol{\mu}}\, s]_{\boldsymbol{\mu}=\boldsymbol{\mu}_{\text{old}},\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_{\text{old}}})^T. \end{aligned}$$

It follows that (6) is equivalent to the updates:

$$\left\{ \begin{array}{ll} \boldsymbol{\Sigma} \leftarrow & \left\{ -2\,\text{vec}^{-1}((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\, s)^T) \right\}^{-1} \\ \boldsymbol{\mu} \leftarrow & \boldsymbol{\mu} + \boldsymbol{\Sigma}\,(\mathsf{D}_{\boldsymbol{\mu}}\, s)^T. \end{array} \right. \tag{7}$$

The simplified form of $\boldsymbol{V}^{-1}\boldsymbol{U}$ in Theorem 1 can be explained via the inverse relationship that exists between $\boldsymbol{V} = \mathsf{H}_{\boldsymbol{\eta}} A(\boldsymbol{\eta})$ and the derivative of the *mean parameter* vector $E\{\boldsymbol{T}(\boldsymbol{x})\}$ with respect to the natural parameter vector $\boldsymbol{\eta}$. This relationship is pointed out in Section 4.1 of Hensman et al. (2012). Note that my $\boldsymbol{U}$ involves the derivative of $[\boldsymbol{\mu}^T\ \text{vec}(\boldsymbol{\Sigma})^T]^T$, rather than $E\{\boldsymbol{T}(\boldsymbol{x})\}$, with respect to $\boldsymbol{\eta}$ in the chain rule. This corresponds to differentiation of s with respect to the more convenient $\text{vec}(\boldsymbol{\Sigma})$.

The update for $\boldsymbol{\Sigma}$, given at (7), involves $\text{vec}^{-1}((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma})}\, s)^T)$. Simplification of this expression for regression models is aided by:

Theorem 2 *Let $\boldsymbol{A}$ be an $n \times d$ matrix, $\boldsymbol{B}$ be a $d \times d$ matrix and $\boldsymbol{b}$ be an $n \times 1$ vector. Define*

$$\mathcal{Q}(\boldsymbol{A}) \equiv (\boldsymbol{A} \otimes \boldsymbol{1}^T) \odot (\boldsymbol{1}^T \otimes \boldsymbol{A})$$

where $\boldsymbol{1}$ is the $d \times 1$ vector with all entries equal to 1. Then

$$(a) \qquad \text{diagonal}\,(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T) = \mathcal{Q}(\boldsymbol{A})\,\text{vec}(\boldsymbol{B})$$

and

$$(b) \qquad \text{vec}\,(\boldsymbol{A}^T \text{diag}(\boldsymbol{b})\boldsymbol{A}) = \mathcal{Q}(\boldsymbol{A})^T\,\boldsymbol{b}.$$

See Appendix B for a proof of Theorem 2.

The following section illustrates the usefulness of Theorems 1 and 2 for assembling non-conjugate variational message passing algorithms involving Multivariate Normal factors.

5. Illustrations

We now provide illustrations of non-conjugate variational message passing that use Multivariate Normal updates. The first Illustration involves a Poisson mixed model and simulated data. We show, in detail, how Theorems 1 and 2 lead to a simple variational algorithm for such models. The second illustration involves heteroscedastic additive model analysis of data from an air pollution study, using non-conjugate variational message passing methodology with Multivariate Normal factors, recently developed by Menictas and Wand (2014).

5.1 Poisson Mixed Model

Consider the single variance component Poisson mixed model:

$$\begin{aligned} &y_i \,|\, \boldsymbol{\beta}, \boldsymbol{u} \quad \text{independently distributed as} \quad \text{Poisson}[\exp\{(\boldsymbol{X}\,\boldsymbol{\beta} + \boldsymbol{Z}\,\boldsymbol{u})_i\}], \quad 1 \le i \le n, \\ &\boldsymbol{u} \,|\, \boldsymbol{\sigma}^2 \sim N(\boldsymbol{0}, \sigma^2\,\boldsymbol{I}_K), \quad \sigma \sim \text{Half-Cauchy}(A) \quad \text{and} \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_p), \end{aligned} \tag{8}$$

where $\boldsymbol{X}$ is a $n \times p$ fixed effects design matrix, $\boldsymbol{Z}$ is a $n \times K$ random effects design matrix and $\sigma \sim \text{Half-Cauchy}(A)$ means that

$$p(\sigma) = \frac{2}{\pi A\{1 + (\sigma/A)^2\}}, \qquad \sigma > 0.$$

Note that, courtesy of Result 5 of Wand et al. (2011), one can replace $\sigma \sim \text{Half-Cauchy}(A)$ by the more convenient auxiliary variable representation

$$\sigma^2 |\, a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a), \qquad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/A^2),$$

where $v \sim \text{Inverse-Gamma}(A, B)$ means that

$$p(v) = \frac{B^A}{\Gamma(A)}\, v^{-A-1}\, \exp(-B/v), \qquad v > 0.$$

Consider the mean field approximation

$$p(\sigma^2, a, \boldsymbol{\beta}, \boldsymbol{u}, |\boldsymbol{y}) \approx q(\sigma^2)\, q(a)\, q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \tag{9}$$

where

$$q(\boldsymbol{\beta}, \boldsymbol{u}; \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \quad \text{is the} \quad N(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \quad \text{density function.}$$

Then application of (2) leads to the optimal q-densities for σ^2 and a being such that

$$\begin{aligned} &q^*(\sigma^2) \text{ is an Inverse-Gamma}(\tfrac{1}{2}(K+1), B_{q(\sigma^2)}) \text{ density function} \\ \text{and} \quad &q^*(a) \text{ is an Inverse-Gamma}(1, B_{q(a)}) \text{ density function} \end{aligned}$$

for rate parameters $B_{q(\sigma^2)}$ and $B_{q(a)}$. Let

$$\mu_{q(1/\sigma^2)} = E_{q(\sigma^2)}(1/\sigma^2) = \tfrac{1}{2}(K+1)/B_{q(\sigma^2)}$$

and $\mu_{q(1/a)}$ be defined similarly. Also let $\boldsymbol{\mu}_{q(\boldsymbol{u})}$ and $\boldsymbol{\Sigma}_{q(\boldsymbol{u})}$ be mean vector and covariance matrix of $q^*(\boldsymbol{u})$. Lastly, let

$$\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}].$$

Algorithm 1 provides explicit forms of the updates required to obtain the optimal parameters of $q^*(\boldsymbol{\beta}, \boldsymbol{u})$, $q^*(a)$ and $q^*(\sigma^2)$.

The derivation of Algorithm 1 is given in Appendix C. The approximate marginal log-likelihood admits the explicit expression:

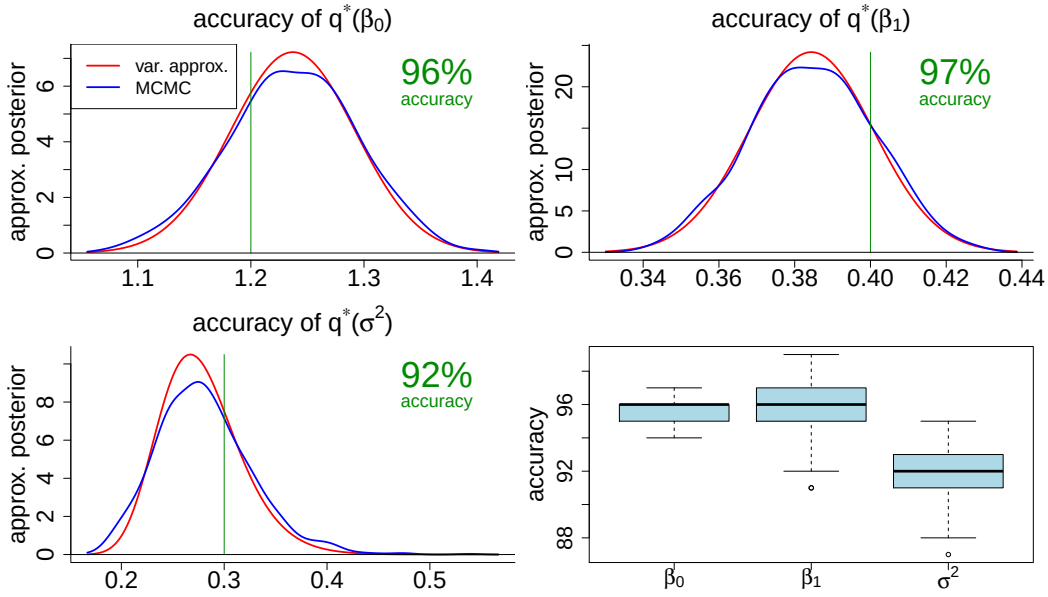

Figure 1: Upper panels and lower left panel: approximate posterior density functions for β_0, β_1 and σ^2 based on the variational approximation scheme described by Algorithm 1 and MCMC, for the first replication of the simulation study described in the text. Accuracy values, according to (11), with the exact posterior density function replaced by the MCMC-based posterior density function are also given. Lower right panel: Side-by-side boxplots of all 1000 accuracy values obtained for each parameter in the simulation study.

$$\begin{aligned}\log \underline{p}(\boldsymbol{y};q) \;=\; & \tfrac{1}{2}(K+p)+\log\Gamma(\tfrac{1}{2}(K+1))-\log(\pi)-\log(A)-\mathbf{1}^T\log(\boldsymbol{y}!)-\tfrac{1}{2}p\log(\sigma^2_{\boldsymbol{\beta}}) \\ & +\boldsymbol{y}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}-\mathbf{1}^T\exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}+\tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\right\} \\ & -\tfrac{1}{2\sigma^2_{\boldsymbol{\beta}}}\{\|\boldsymbol{\mu}_{q(\boldsymbol{\beta})}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})})\}+\tfrac{1}{2}\log|\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}| \\ & -\tfrac{1}{2}(K+1)\log\left(\tfrac{1}{2}\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2+\text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u})})\}+\mu_{q(1/a)}\right) \\ & -\log(\mu_{q(1/\sigma^2)}+A^{-2})+\mu_{q(1/\sigma^2)}\mu_{q(1/a)}.\end{aligned}$$

It is noteworthy that the variational message passing algorithm with derived variables, as described in Appendix A of Minka and Winn (2008), leads to an alternative to Algorithm 1 that requires only Univariate Normal updates corresponding to (7) of Knowles and Minka (2011). Such an approach is used in the Infer.NET computational framework (Minka et al., 2013). However, this alternative version is not as succinct as Algorithm 1. The simplified version that arises from Theorems 1 and 2 allows easier extension to more complicated models.

Initialize: $\mu_{q(1/\sigma^2)} > 0$, $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ a $(p+K)\times 1$ vector and $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ a $(p+K)\times(p+K)$ positive definite matrix.
Cycle:

$$\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \exp\left\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\right\}$$

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \left(\boldsymbol{C}^T\text{diag}\{\boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}\}\,\boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\right)^{-1}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\left\{\boldsymbol{C}^T(\boldsymbol{y} - \boldsymbol{w}_{q(\boldsymbol{\beta},\boldsymbol{u})}) - \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\right\}$$

$$\mu_{q(1/\sigma^2)} \leftarrow \frac{K+1}{2\mu_{q(1/a)} + \|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \text{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u})})} \quad ; \quad \mu_{q(1/a)} \leftarrow 1/(\mu_{q(1/\sigma^2)} + A^{-2}).$$

until the absolute change in $\underline{p}(\boldsymbol{y};q)$ is negligible.

Algorithm 1: Iterative scheme for determination of the optimal parameters in $q^*(\boldsymbol{\beta},\boldsymbol{u})$, $q^*(\sigma^2)$ and $q^*(a)$ for the posterior density function approximation (9).

I replicated 1000 data-sets corresponding to the simulation setting

$$\begin{gathered} y_{ij}|U_i \sim \text{Poisson}\left(\exp(\beta_0 + \beta_1\, x_{ij} + U_i)\right),\ U_i|\sigma^2 \sim N(0,\sigma^2), \\ 1 \le i \le m,\ 1 \le j \le n, \\ \sigma^2|a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, 1/a), \quad a \sim \text{Inverse-Gamma}(\tfrac{1}{2}, A^{-2}), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\boldsymbol{\beta}}^2\,\boldsymbol{I}) \end{gathered} \tag{10}$$

The hyperparameters were set at $\sigma_{\boldsymbol{\beta}} = A = 10^5$ and the sample sizes were $m = 100$, $n = 10$. Note that (10) is a special case of (8) with $\boldsymbol{Z} = \boldsymbol{I}_m \otimes \boldsymbol{1}_n$, where $\boldsymbol{1}_n$ is the $n \times 1$ vector with all entries equal to one.

For each data-set I obtained approximate posterior density functions for β_0, β_1 and σ^2 using both Algorithm 1 and Markov chain Monte Carlo (MCMC). For MCMC I used the package `BRugs` (Ligges et al., 2012) within the R computing environment (R Development Core Team, 2013) with a burnin of size 5000 followed by the generation of 5000 samples, with a thinning factor of 5. This resulted in MCMC samples of size 1000 being retained for inference. The iterations in Algorithm 1 were terminated when the relative change in $\log \underline{p}(\boldsymbol{y};q)$ fell below 10^{-4}.

Figure 1 displays side-by-side boxplots of accuracy scores defined by

$$\text{accuracy}(q^*) = 100\left(1 - \tfrac{1}{2}\int_{-\infty}^{\infty} \left|\, q^*(\theta) - p(\theta|\boldsymbol{y}) \,\right| d\theta\right)\%. \tag{11}$$

for a generic parameter θ, and with $p(\theta|\boldsymbol{y})$ replaced by a kernel density estimate based on the MCMC sample. The boxplots show that the majority of accuracy scores exceed 95%, and that they rarely drop below 90%.

Figure 1 allows visual assessment of the variational approximate posterior density functions against the MCMC-based benchmark for a single replication of the simulation study. The accuracy is seen to be excellent for β_0 and β_1 and very good for σ^2.

As discussed in Knowles and Minka (2011), convergence of non-conjugate variational message passing is not guaranteed. In the simulation study the algorithm converged in all replications regardless of starting values, but in about 2% of the cases this required some adjustment to avoid inverting a singular matrix in the $\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ update during the early iterations. The adjustment involves adding $\varepsilon \boldsymbol{I}$ to the matrix requiring inversion, with $\varepsilon > 0$ chosen so that the condition number stayed below 10^{16}. In almost all cases, this adjustment was only necessary for the first few iterations.

In this section we have shown that non-conjugate variational message passing leads to an attractive variational inference algorithm for Poisson mixed models. Since the exponential moments of Multivariate Normal random vectors are available in closed form, no quadrature is required in the Poisson case. Other generalized linear mixed models, such as logistic mixed models, require quadrature. The logistic analogue of (8) is such that only univariate quadrature is required. Details are given in Appendix B of Tan and Nott (2013).

5.2 Heteroscedastic Additive Model

This illustration involves analysis of data from the Californian air pollution study described in Breiman and Friedman (1985). The response variable is

$$y = \text{ozone concentration (ppm) at Sandburg Air Force Base}$$

and three predictors variables are

$$\begin{aligned} x_1 &= \text{pressure gradient (mm Hg) from Los Angeles International Airport to Daggett, California,} \\ x_2 &= \text{inversion base height (feet)} \\ \text{and } x_3 &= \text{inversion base temperature (degrees Fahrenheit).} \end{aligned}$$

The data comprises 345 measurements on each of these 4 variables. Let $(x_{i1}, x_{i3}, x_{i3}, y_i)$, $1 \leq i \leq 345$ denote the full regression data set.

We entertained the *heteroscedastic additive model*

$$y_i \sim N\Big(\beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}), \exp\big(\gamma_0 + h_1(x_{1i}) + h_2(x_{2i}) + h_3(x_{3i})\big)\Big), \qquad (12)$$

for $1 \leq i \leq 345$. Here f_j and g_j, $j = 1, 2, 3$, and smooth but otherwise arbitrary functions. Bayesian mixed model-based penalized splines (e.g. Ruppert et al., 2003) were used to model the smooth functions as follows:

$$\begin{aligned} f_j(x) &= \beta_j\, x + \sum_{k=1}^{K_j} u_{jk}\, z_{jk}(x), \quad u_{jk} \text{ iid } N(0, \sigma_{uj}^2) \\ \text{and} \quad h_j(x) &= \gamma_j\, x + \sum_{k=1}^{K_j} v_{jk}\, z_{jk}(x), \quad v_{jk} \text{ iid } N(0, \sigma_{vj}^2). \end{aligned} \qquad (13)$$

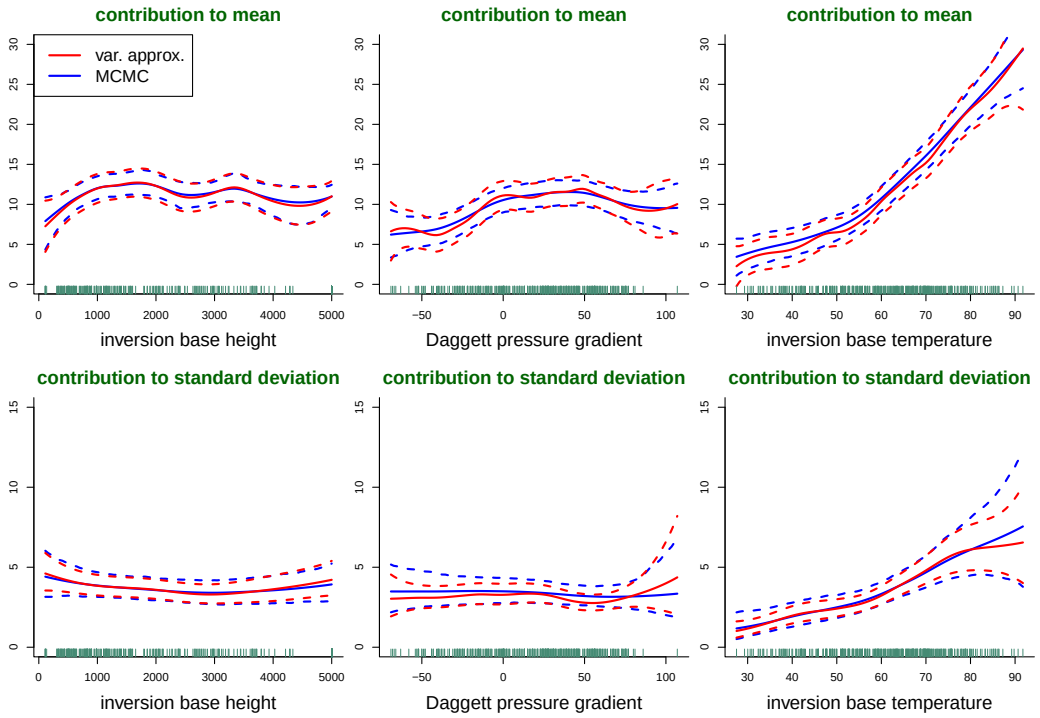

Figure 2: Upper panels: approximate pointwise posterior means and 95% credible sets for the mean function contributions f_1, f_2 and f_3 according to the heteroscedastic additive model (12). Vertical alignment of the estimated functions is described in the text. Lower panels: approximate pointwise posterior means and 95% credible sets for the standard deviation function contributions $\exp(h_2/2)$, $\exp(h_2/2)$ and $\exp(h_3/2)$. Approximate Bayesian inference is based on both non-conjugate variational message passing and MCMC.

where iid stands for 'independently and identically distributed as'. The $\{z_{jk} : 1 \le k \le K_j\}$, $j = 1, 2, 3$, are spline bases of sizes K_j respectively. My default for the z_{jk} are suitably transformed cubic O'Sullivan splines, as described in Section 4 of Wand and Ormerod (2008). The priors on the regression coefficients and standard deviation parameters are

$$\beta_j \text{ iid } N(0, \sigma^2_{\boldsymbol{\beta}}),\ \gamma_j \text{ iid } N(0, \sigma^2_{\boldsymbol{\gamma}}),\ \sigma_{uj} \text{ iid Half-Cauchy}(A_u),\ \sigma_{vj} \text{ iid Half-Cauchy}(A_v). \quad (14)$$

The regression data replaced by standardized versions and the hyperparameters were set to be $\sigma_{\boldsymbol{\beta}} = \sigma_{\boldsymbol{\gamma}} = A_u = A_v = 10^5$, corresponding to non-informativity. The results were transformed to the original units after fitting. The basis function sizes were all fixed at $K_1 = K_2 = K_3 = 18$.

The Bayesian model given by (12), (13) and (14) admits a closed form non-conjugate variational message passing algorithm, with the regression coefficients for the full mean and variance functions each being Multivariate Normal. Details are given in Menictas and Wand (2014). Figure 2 shows the estimated mean function (f_j) contributions, and the standard deviation function ($\exp(h_j/2)$) contributions based on both variational approximation

and MCMC. The MCMC inference was carried out in the same fashion as for the illustration described in Section 5.1. The abbreviated names `inversion base height`, `Daggett pressure gradient` and `inversion base temperature` are used for x_1, x_2 and x_3. The estimated f_1 display is vertically aligned to match the response data by evaluating the estimate of f_2 at $\overline{x}_2$ and estimate of f_3 at $\overline{x}_3$, where $\overline{x}_2$ and $\overline{x}_3$ are the sample means of the x_{2i} and x_{3i}, respectively. Analogous alignment strategies were used for the f_2 and f_3 displays.

Figure 2 shows that there is excellent agreement between non-conjugate variational message passing, with Multivariate Normal coefficient vectors, and MCMC. The former approach is considerably faster. The heteroscedasticity is seen to be relatively mild for `inversion base height` and `Daggett pressure gradient`. However, there is pronounced heteroscedasticity in `inversion base temperature` that is captured by model (12).

Acknowledgments

This research was partially supported by Australian Research Council Discovery Project DP110100061. The author is grateful to Cathy Lee, Jan Luts, Marianne Menictas, Tom Minka, David Nott and Linda Tan for their comments.

Appendix A: Proof of Theorem 1

Proof of (a)

For the upper-left block of $\boldsymbol{U}$ note that $\boldsymbol{\mu} = \boldsymbol{\Sigma}\,\boldsymbol{\eta}_1$ and so $d_{\boldsymbol{\eta}_1}\boldsymbol{\mu} = \boldsymbol{\Sigma}\,d\boldsymbol{\eta}_1$. Theorem 6 of Magnus and Neudecker (1999) leads to $\mathsf{D}_{\boldsymbol{\eta}_1}\boldsymbol{\mu} = \boldsymbol{\Sigma}$. The lower-right block of $\boldsymbol{U}$ involves the relation $\boldsymbol{\Sigma} = -\frac{1}{2}\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\boldsymbol{\eta}_2)\}^{-1}$ given in (5). Then Rule 3.3.5 in Wand (2002) and the identity

$$\mathrm{vec}(\boldsymbol{ABC}) = (\boldsymbol{C}^T \otimes \boldsymbol{A})\,\mathrm{vec}(\boldsymbol{B}) \tag{15}$$

leads to

$$d_{\boldsymbol{\eta}_2}\,\mathrm{vec}(\boldsymbol{\Sigma}) = 2\,\mathrm{vec}(\boldsymbol{\Sigma}\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\,d\boldsymbol{\eta}_2)\}\,\boldsymbol{\Sigma}) = 2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\,\boldsymbol{D}_d^{+T}\,d\boldsymbol{\eta}_2.$$

Hence, making use of Theorem 13 (b), Chapter 3, of Magnus and Neudecker (1999),

$$\mathsf{D}_{\boldsymbol{\eta}_2}\,\mathrm{vec}(\boldsymbol{\Sigma}) = 2(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\,\boldsymbol{D}_d^{+T} = 2\boldsymbol{D}_d\boldsymbol{D}_d^{+}(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma})\,\boldsymbol{D}_d^{+T} = \boldsymbol{D}_d\boldsymbol{S} = (\boldsymbol{S}\boldsymbol{D}_d^{T})^T.$$

The expression for the lower-left block of $\boldsymbol{U}$ follows from

$$d_{\boldsymbol{\eta}_2}\,\boldsymbol{\mu} = 2\boldsymbol{\Sigma}\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\,d\boldsymbol{\eta}_2)\}\,\boldsymbol{\Sigma}\boldsymbol{\eta}_1 = 2\,\mathrm{vec}(\boldsymbol{\Sigma}\{\mathrm{vec}^{-1}(\boldsymbol{D}_d^{+T}\,d\boldsymbol{\eta}_2)\}\,\boldsymbol{\mu}) = 2(\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}\,d\boldsymbol{\eta}_2$$

where Rule 3.3.5 in Wand (2002) and (15) have been used again. This gives

$$\mathsf{D}_{\boldsymbol{\eta}_2}\,\boldsymbol{\mu} = 2(\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T} = \{2\boldsymbol{D}_d^{+}(\boldsymbol{\mu}\otimes\boldsymbol{\Sigma})\}^T = \{2\boldsymbol{D}_d^{+}(\boldsymbol{\mu}\otimes\boldsymbol{I}_d)(1\otimes\boldsymbol{\Sigma})\}^T = (\boldsymbol{M}\boldsymbol{\Sigma})^T.$$

For the upper-left block note that, from (5), $d_{\boldsymbol{\eta}_1}\mathrm{vec}(\boldsymbol{\Sigma}) = \boldsymbol{0}\,d_{\boldsymbol{\eta}_1}$ and so $\mathsf{D}_{\boldsymbol{\eta}_1}\mathrm{vec}(\boldsymbol{\Sigma}) = \boldsymbol{0}$.

Proof of (b)

This is an immediate consequence of Theorem 13(d), Chapter 3, of Magnus and Neudecker (1999).

Proof of (c)

The upper-left block is $\text{var}(\boldsymbol{x}) = \boldsymbol{\Sigma}$. The lower-right block is

$$\begin{aligned}\text{var}\{\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\} &= \boldsymbol{D}_d^+ \,\text{var}\{\text{vec}(\boldsymbol{x}1\boldsymbol{x}^T)\}\boldsymbol{D}_d^{+T} = \boldsymbol{D}_d^+ \,\text{var}(\text{vec}(\boldsymbol{x}\otimes\boldsymbol{x}^T))\boldsymbol{D}_d^{+T} \\ &= \boldsymbol{D}_d^+ \,(\boldsymbol{I}_{d^2} + \boldsymbol{K}_d)(\boldsymbol{\Sigma}\otimes\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\otimes\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}\end{aligned}$$

where (15) and Theorem 4.3 (iv) of Magnus and Neudecker (1979) has been used. Here $\boldsymbol{K}_d$ denotes the *commutation matrix* of order d, defined by $\boldsymbol{K}_d(\boldsymbol{A}\otimes\boldsymbol{B}) = (\boldsymbol{B}\otimes\boldsymbol{A})\boldsymbol{K}_d$ for arbitrary $d\times d$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Noting the identity $\frac{1}{2}\,\boldsymbol{D}_d^+(\boldsymbol{I}_{d^2}+\boldsymbol{K}_d) = \boldsymbol{D}_d^+$, which is an immediate consequence of (15) in Chapter 3 of Magnus and Neudecker (1999), one then gets

$$\text{var}\{\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\} = \boldsymbol{S} + 2\boldsymbol{D}_d^+(\boldsymbol{\Sigma}\otimes\boldsymbol{\mu}\boldsymbol{\mu}^T + \boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}.$$

Theorem 12(a), Chapter 3, of Magnus and Neudecker (1999) states that $\boldsymbol{K}_d\boldsymbol{D}_d = \boldsymbol{D}_d$, which implies that

$$\boldsymbol{D}_d^+ = \{(\boldsymbol{K}_d\boldsymbol{D}_d)^T\boldsymbol{K}_d\boldsymbol{D}_d\}^{-1}\boldsymbol{D}_d^T\boldsymbol{K}_d^T = (\boldsymbol{D}_d^T\boldsymbol{K}_d^T\boldsymbol{K}_d\boldsymbol{D}_d)^{-1}\boldsymbol{D}_d^T\boldsymbol{K}_d = \boldsymbol{D}_d^+\boldsymbol{K}_d. \qquad (16)$$

Here I have used $\boldsymbol{K}^T = \boldsymbol{K}_d^{-1} = \boldsymbol{K}_d$ as stated in (2) of Chapter 3 of Magnus and Neudecker (1999). The identity $\boldsymbol{D}_d^+ = \boldsymbol{D}_d^+\boldsymbol{K}_d$ leads to

$$\boldsymbol{D}_d^+(\boldsymbol{\Sigma}\otimes\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{D}_d^{+T} = \boldsymbol{D}_d^+\boldsymbol{K}_d(\boldsymbol{\Sigma}\otimes\boldsymbol{\mu}\boldsymbol{\mu}^T)\boldsymbol{D}_d^{+T} = \boldsymbol{D}_d^+(\boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{K}_d^T\boldsymbol{D}_d^{+T} = \boldsymbol{D}_d^+(\boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}$$

leading to $\text{var}\{\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\} = \boldsymbol{S} + 4\boldsymbol{D}_d^+(\boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma})\boldsymbol{D}_d^{+T}$. Since

$$(\boldsymbol{\mu}\boldsymbol{\mu}^T\otimes\boldsymbol{\Sigma}) = (\boldsymbol{\mu}\otimes\boldsymbol{\Sigma})(\boldsymbol{\mu}^T\otimes\boldsymbol{I}_d) = (\boldsymbol{\mu}\otimes\boldsymbol{I}_d)(1\otimes\boldsymbol{\Sigma})(\boldsymbol{\mu}^T\otimes\boldsymbol{I}_d) = (\boldsymbol{\mu}\otimes\boldsymbol{I}_d)\boldsymbol{\Sigma}(\boldsymbol{\mu}\otimes\boldsymbol{I}_d)^T.$$

I conclude that

$$\text{var}\{\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)\} = \boldsymbol{S} + \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T.$$

The (i,j) entry of the lower-left block is

$$\text{cov}(\text{vech}(\boldsymbol{x}\boldsymbol{x}^T)_i, x_j) = \text{cov}(\{\boldsymbol{D}_d^+\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)\}_i, x_j) = \sum_{k=1}^{d^2}(\boldsymbol{D}_d^+)_{ik}\,\text{cov}(\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)_k, x_j). \qquad (17)$$

Let $\lfloor x\rfloor$ denote the largest integer less than or equal to x. Then using one of the fundamental identities for generalized cumulants given on page 58 of McCullagh (1987),

$$\begin{aligned}\text{cov}(\text{vec}(\boldsymbol{x}\boldsymbol{x}^T)_k, x_j) &= \text{cov}(x_{k-d\lfloor(k-1)/d\rfloor}\, x_{\lfloor(k-1)/d\rfloor+1}\,,\, x_j) \\ &= \mu_{k-d\lfloor(k-1)/d\rfloor}\Sigma_{\lfloor(k-1)/d\rfloor+1,j} + \mu_{\lfloor(k-1)/d\rfloor+1}\Sigma_{k-d\lfloor(k-1)/d\rfloor,j} \\ &= (\boldsymbol{\Sigma}\otimes\boldsymbol{\mu})_{kj} + (\boldsymbol{\mu}\otimes\boldsymbol{\Sigma})_{kj}.\end{aligned}$$

Combining this with (17), the lower-left block equals $\boldsymbol{D}_d^+(\boldsymbol{\Sigma}\otimes\boldsymbol{\mu} + \boldsymbol{\mu}\otimes\boldsymbol{\Sigma})$. But, courtesy of (16), this equals

$$\boldsymbol{D}_d^+(\boldsymbol{\mu}\otimes\boldsymbol{\Sigma}) + \boldsymbol{D}_d^+\boldsymbol{K}_d(\boldsymbol{\Sigma}\otimes\boldsymbol{\mu}) = 2\boldsymbol{D}_d^+(\boldsymbol{\mu}\otimes\boldsymbol{\Sigma}) = \boldsymbol{M}\boldsymbol{\Sigma}.$$

Proof of (d)

It is straightforward to verify that

$$\begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma}\boldsymbol{M}^T \\ \boldsymbol{M}\boldsymbol{\Sigma} & \boldsymbol{S} + \boldsymbol{M}\boldsymbol{\Sigma}\boldsymbol{M}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}^{-1} + \boldsymbol{M}^T\boldsymbol{S}^{-1}\boldsymbol{M} & -\boldsymbol{M}^T\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M} & \boldsymbol{S}^{-1} \end{bmatrix} = \boldsymbol{I}_{d+d(d+1)/2}.$$

The stated expression for $\boldsymbol{V}^{-1}$ immediately follows.

Proof of (e)

$$\boldsymbol{V}^{-1}\boldsymbol{U} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} + \boldsymbol{M}^T\boldsymbol{S}^{-1}\boldsymbol{M} & -\boldsymbol{M}^T\boldsymbol{S}^{-1} \\ -\boldsymbol{S}^{-1}\boldsymbol{M} & \boldsymbol{S}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{M}\boldsymbol{\Sigma} & \boldsymbol{S}\boldsymbol{D}_d^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{M}^T\boldsymbol{D}_d^T \\ \boldsymbol{0} & \boldsymbol{D}_d^T \end{bmatrix}.$$

Proof of (f)

First note that

$$\boldsymbol{V}^{-1}\boldsymbol{U} \begin{bmatrix} \boldsymbol{g} \\ \mathrm{vec}(\boldsymbol{G}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{M}^T\boldsymbol{D}_d^T \\ \boldsymbol{0} & \boldsymbol{D}_d^T \end{bmatrix} \begin{bmatrix} \boldsymbol{g} \\ \mathrm{vec}(\boldsymbol{G}) \end{bmatrix} = \begin{bmatrix} \boldsymbol{g} - \boldsymbol{M}^T\boldsymbol{D}_d^T\,\mathrm{vec}(\boldsymbol{G}) \\ \boldsymbol{D}_d^T\mathrm{vec}(\boldsymbol{G}) \end{bmatrix}.$$

With the help of Lemma 1 and (15) one then has

$$\boldsymbol{M}^T\boldsymbol{D}_d^T\,\mathrm{vec}(\boldsymbol{G}) = 2(\boldsymbol{\mu}^T \otimes \boldsymbol{I}_d)\boldsymbol{D}_d^{+T}\boldsymbol{D}_d^T\mathrm{vec}(\boldsymbol{G}) = 2(\boldsymbol{\mu}^T \otimes \boldsymbol{I}_d)\,\mathrm{vec}(\boldsymbol{G}) = 2\,\boldsymbol{G}\boldsymbol{\mu}$$

and the stated result is obtained.

Appendix B: Proof of Theorem 2

Proof of (a)

Let A_{ij} and B_{ij}, respectively, denote the (i,j) entry of $\boldsymbol{A}$ and $\boldsymbol{B}$. Then a listing of the entries of $\mathcal{Q}(\boldsymbol{A})$ reveals that its entry is (i,j) is

$$\mathcal{Q}(\boldsymbol{A})_{ij} = A_{i,\lfloor (j-1)/d \rfloor + 1}\, A_{i,j-d\lfloor (j-1)/d \rfloor}, \quad 1 \le i \le n,\ 1 \le j \le d^2. \tag{18}$$

Similarly, the ith entry of $\mathrm{vec}(\boldsymbol{B})$ is

$$\mathrm{vec}(\boldsymbol{B})_j = B_{j-d\lfloor (j-1)/d \rfloor, \lfloor (j-1)/d \rfloor + 1}, \quad 1 \le j \le d^2. \tag{19}$$

Hence

$$\begin{aligned}
\{\mathcal{Q}(\boldsymbol{A})\,\mathrm{vec}(\boldsymbol{B})\}_i &= \sum_{j=1}^{d^2} A_{i,\lfloor (j-1)/d \rfloor + 1}\, A_{i,j-d\lfloor (j-1)/d \rfloor}\, B_{j-d\lfloor (j-1)/d \rfloor, \lfloor (j-1)/d \rfloor + 1} \\
&= \left(\sum_{j=1}^{d} + \sum_{j=d+1}^{2d} + \ldots + \sum_{j=(d-1)\,d+1}^{d^2} \right) A_{i,\lfloor (j-1)/d \rfloor + 1}\, A_{i,j-d\lfloor (j-1)/d \rfloor} \\
&\qquad \times B_{j-d\lfloor (j-1)/d \rfloor, \lfloor (j-1)/d \rfloor + 1} \\
&= \sum_{j=1}^{d} A_{i1}A_{ij}B_{1j} + \sum_{j=1}^{d} A_{i2}A_{ij}B_{2j} + \ldots + \sum_{j=1}^{d} A_{id}A_{ij}B_{dj} \\
&= \sum_{j=1}^{d}\sum_{j'=1}^{d} A_{ij}A_{ij'}B_{jj'} = \mathrm{diagonal}(\boldsymbol{A}\boldsymbol{B}\boldsymbol{A}^T)_i
\end{aligned}$$

and the result follows immediately.

Proof of (b)

Letting b_i denote the ith entry of $\boldsymbol{b}$ and making use of (18) one has

$$\begin{aligned}\{\mathcal{Q}(\boldsymbol{A})^T\,\boldsymbol{b}\}_j &= \sum_{i=1}^{n}\{\mathcal{Q}(\boldsymbol{A})^T\}_{ji}\,b_i = \sum_{i=1}^{n}\mathcal{Q}(\boldsymbol{A})_{ij}\,b_i \\ &= \sum_{i=1}^{n} b_i\,A_{i,\lfloor(j-1)/d\rfloor+1}\,A_{i,j-d\lfloor(j-1)/d\rfloor}.\end{aligned}$$

Application of (19) to $\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{b})\boldsymbol{A}$ gives

$$\begin{aligned}\mathrm{vec}(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{b})\boldsymbol{A})_j &= (\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{b})\boldsymbol{A})_{j-d\lfloor(j-1)/d\rfloor,\,\lfloor(j-1)/d\rfloor+1} \\ &= \sum_{i=1}^{n} b_i(\boldsymbol{A}^T)_{j-d\lfloor(j-1)/d\rfloor,i}\boldsymbol{A}_{i,\lfloor(j-1)/d\rfloor+1} \\ &= \sum_{i=1}^{n} b_i\,A_{i,\lfloor(j-1)/d\rfloor+1}\,A_{i,j-d\lfloor(j-1)/d\rfloor} = \{\mathcal{Q}(\boldsymbol{A})^T\,\boldsymbol{b}\}_j\end{aligned}$$

which proves equality between $\mathcal{Q}(\boldsymbol{A})^T\,\boldsymbol{b}$ and $\mathrm{vec}(\boldsymbol{A}^T\mathrm{diag}(\boldsymbol{b})\boldsymbol{A})$.

Appendix C: Derivation of Algorithm 1

Derivation of $q^*(\sigma^2)$

$$\begin{aligned}\log q^*(\sigma^2) &= E_q\{\log p(\sigma^2|\text{rest})\} + \text{const} \\ &= \{-\tfrac{1}{2}(K+1)-1\}\log(\sigma^2) - \{\tfrac{1}{2}E_q\|\boldsymbol{u}\|^2 + \mu_{q(1/a)}\}/\sigma^2 + \text{const.}\end{aligned}$$

where 'const' denotes terms not involving σ^2. Using

$$E_q\|\boldsymbol{u}\|^2 = \|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\mu})})$$

I then get $q^*(\sigma^2) \sim \text{Inverse-Gamma}(\frac{1}{2}(K+1), B_{q(\sigma^2)})$ where

$$B_{q(\sigma^2)} = \tfrac{1}{2}\{\|\boldsymbol{\mu}_{q(\boldsymbol{u})}\|^2 + \mathrm{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{\mu})})\} + \mu_{q(1/a)}.$$

Derivation of $q^*(a)$

$$\begin{aligned}\log q^*(a) &= E_q\{\log p(a|\text{rest})\} + \text{const} \\ &= (-1-1)\log(a) - (\mu_{q(1/\sigma^2)} + A^{-2})/a + \text{const.}\end{aligned}$$

This gives $q^*(a) \sim \text{Inverse-Gamma}(1, B_{q(a)})$ where

$$B_{q(a)} = \mu_{q(1/\sigma^2)} + A^{-2}.$$

Derivation of the $(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})$ Updates

Note that

$$\begin{aligned} E_q\{\log p(\boldsymbol{y},\boldsymbol{\beta},\boldsymbol{u},\sigma^2,a)\} &= E_q\{\log\, p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}) + \log\, p(\boldsymbol{\beta},\boldsymbol{u}|\sigma^2) + \log\, p(\sigma^2|a) + \log\, p(a)\} \\ &= S + \text{terms not involving } \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \text{ or } \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \end{aligned}$$

where

$$S \equiv S(\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \equiv E_q\{\log\, p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u}) + \log\, p(\boldsymbol{\beta},\boldsymbol{u}|\sigma^2)\}.$$

Then

$$\begin{aligned} S &= \boldsymbol{y}^T\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} - \mathbf{1}^T \exp\Big\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\Big\} \\ &\quad -\tfrac{1}{2}\text{tr}\left(\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T + \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\}\right) \\ &\quad -\tfrac{1}{2}(p+K)\log(2\pi) - \tfrac{1}{2}\,p\log(\sigma_{\boldsymbol{\beta}}^2) - \tfrac{1}{2}\,K\,E_q\{\log(\sigma^2)\} - \mathbf{1}^T\log(\boldsymbol{y}!) \end{aligned}$$

and so

$$\begin{aligned} d_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} S &= \boldsymbol{y}^T\boldsymbol{C}\, d\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ &\quad -\mathbf{1}^T\text{diag}[\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\boldsymbol{C}\, d\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ &\quad -\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix} d\,\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ &= \Big(\Big[\boldsymbol{y} - \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}\Big]^T\boldsymbol{C} \\ &\quad -\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}^T\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\Big) d\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}. \end{aligned}$$

Thus, by Theorem 6, Chapter 5, of Magnus and Neudecker (1999),

$$\begin{aligned} \{\mathsf{D}_{\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}} S\}^T &= \boldsymbol{C}^T\Big[\boldsymbol{y} - \exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}\Big] \\ &\quad -\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}. \end{aligned}$$

Next, using Theorem 2 of Section 4 and Rule 3.3.2 of Wand (2002),

$$\begin{aligned} d_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})} S &= -\mathbf{1}^T\text{diag}[\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\tfrac{1}{2}\mathcal{Q}(\boldsymbol{C})\, d\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \\ &\quad -\tfrac{1}{2}\text{vec}\left(\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\right)^T d\,\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \\ &= \Big(-\tfrac{1}{2}\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}^T\mathcal{Q}(\boldsymbol{C}) \\ &\quad -\tfrac{1}{2}\text{vec}\left(\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\right)^T\Big) d\,\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \\ &= -\tfrac{1}{2}\text{vec}\Big(\boldsymbol{C}^T\text{diag}[\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\boldsymbol{C} \\ &\quad +\begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \mathbf{0} \\ \mathbf{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\Big)^T d\,\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}) \end{aligned}$$

and so

$$\text{vec}^{-1}\left((\mathsf{D}_{\text{vec}(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})})}\,S)^T\right) = -\tfrac{1}{2}\left(\boldsymbol{C}^T\text{diag}[\exp\{\boldsymbol{C}\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})} + \tfrac{1}{2}\text{diagonal}(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^T)\}]\,\boldsymbol{C} + \begin{bmatrix} \sigma_{\boldsymbol{\beta}}^{-2}\boldsymbol{I}_p & \boldsymbol{0} \\ \boldsymbol{0} & \mu_{q(1/\sigma^2)}\boldsymbol{I}_K \end{bmatrix}\right).$$

References

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

L. Breiman and J. Friedman. Estimating optimal transformations for multiple regression and correlation (with discussion). *Journal of the American Statistical Association*, 80: 580–619, 1985.

J. Hensman, M. Rattray, and N. D. Lawrence. Fast variational inference in the conjugate exponential family. In P. Bartlett, F.C.N. Pereira, C.J.C. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 2897–2905, 2012.

D. A. Knowles and T. P. Minka. Non-conjugate message passing for multinomial and binary regression. In J. Shawe-Taylor, R.S. Zamel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 1701–1709, 2011.

U. Ligges, S. Sturtz, A. Gelman, G. Gorjanc, and C. Jackson. *BRugs. Fully-interactive R interface to the OpenBUGS software for Bayesian analysis using MCMC sampling*, 2012. URL `http://cran.r-project.org`. R package version 0.8-0.

J. R. Magnus and H. Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics, Revised Edition*. Wiley, Chichester UK, 1999.

J.R. Magnus and H. Neudecker. The commutation matrix: some properties and applications. *The Annals of Statistics*, 7:381–394, 1979.

P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall, London, 1987.

M. Menictas and M. P. Wand. Variational inference for heteroscedastic semiparametric regression. 2014. Unpublished manuscript.

T. Minka and J. Winn. Gates: A graphical notation for mixture models. *Microsoft Research Technical Report Series*, MSR-TR-2008-185:1–16, 2008.

T. Minka, J. Winn, J. Guiver, and D. Knowles. *Infer.NET 2.5*, 2013. URL `http://research.microsoft.com/infernet`. Microsoft Research Cambridge.

F. Novomestky. *matrixcalc. A collection of functions to support matrix differential calculus as presented in Magnus and Neudecker (1999)*, 2008. URL `http://cran.r-project.org`. R package version 1.0-1.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric Regression*. Cambridge University Press, New York, 2003.

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

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

M. P. Wand. Vector differential calculus in statistics. *The American Statistician*, 56:55–62, 2002.

M. P. Wand and J. T. Ormerod. On semiparametric regression with O'Sullivan penalized splines. *Australian and New Zealand Journal of Statistics*, 50:179–198, 2008.

M. P. Wand, J. T. Ormerod, S. A. Padoan, and R. Frühwirth. Mean field variational Bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900, 2011.

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