Australian & New Zealand Journal of Statistics

Aust. N. Z. J. Stat. **53**(3), 2011, 305–330 doi: 10.1111/j.1467-842X.2011.00637.x

GENERALIZED EXTREME VALUE ADDITIVE MODEL ANALYSIS VIA MEAN FIELD VARIATIONAL BAYES

SARAH E. NEVILLE[1], M. J. PALMER[2] AND M. P. WAND[3,*]

University of Wollongong, Commonwealth Scientific and Industrial Research Organisation and University of Technology

Summary

We develop Mean Field Variational Bayes methodology for fast approximate inference in Bayesian Generalized Extreme Value additive model analysis. Such models are useful for flexibly assessing the impact of continuous predictor variables on sample extremes. The new methodology allows large Bayesian models to be fitted and assessed without the significant computing costs of Markov Chain Monte Carlo methods. We illustrate our new methodology with maximum rainfall data from the Sydney, Australia, hinterland. Comparisons are made between the Mean Field Variational Bayes and Markov Chain Monte Carlo approaches.

Key words: auxiliary mixture sampling; Bayesian inference; generalized additive models; sample extremes; variational approximation.

1. Introduction

Regression analysis for sample extreme responses is a topic of considerable current interest, with climate research being one of the main driving forces. The last decade has seen generalized additive models for sample extremes added to the regression armory. Generalized additive models allow for flexible nonlinear relationships between sample extreme responses and continuous predictors. For example, seasonal and spatial effects tend to be highly nonlinear, and penalized splines combined with additivity restrictions have the ability to tease out such effects. Some recent references on generalized additive models for sample extreme responses are Davison & Ramesh (2000), Chavez-Demoulin & Davison (2005), Yee & Stephenson (2007), Padoan & Wand (2008) and Laurini & Pauli (2009), and each of these is motivated by climate research data sets. For example, Chavez-Demoulin & Davison (2005) motivate their methodology using data on minimum daily temperatures at 21 Swiss weather stations for the winters between 1971 and 1997, and their generalized additive model shows that the North Atlantic oscillation index has a nonlinear effect on the parameters of their sample extremes response.

*Author to whom correspondence should be addressed.
[1]Centre for Statistical and Survey Methodology, School of Mathematics and Applied Statistics, University of Wollongong, Wollongong 2522, Australia.
[2]Commonwealth Scientific and Industrial Research Organisation, Division of Mathematics, Informatics and Statistics, Floreat 6014, Australia.
[3]School of Mathematical Sciences, University of Technology, Sydney, Broadway 2007, Australia.
Acknowledgments. The authors are grateful for comments received from a Commonwealth Scientific and Industrial Research Organisation internal reviewer. This research was partially supported by Australian Research Council Discovery Project DP0877055. This work has been undertaken as part of the Australian Climate Change Science Program, funded jointly by Australia's Department of Climate Change and Energy Efficiency, Bureau of Meteorology and Commonwealth Scientific and Industrial Research Organisation.

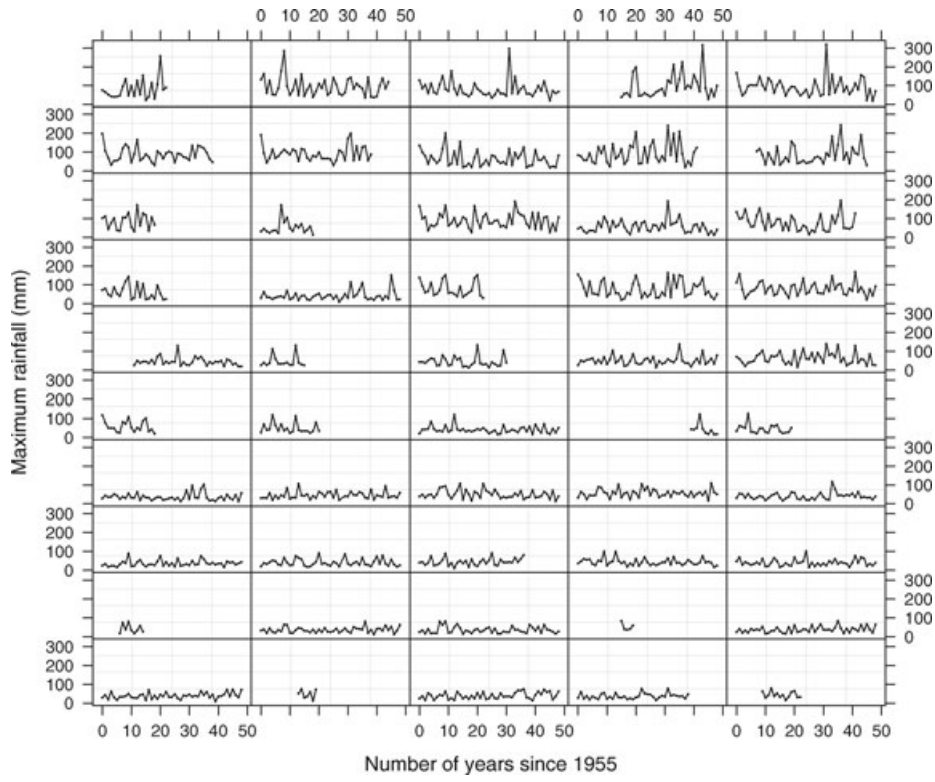

Figure 1. Annual winter maximum rainfall at 50 weather weather stations in the Sydney, Australia, hinterland.

Existing generalized additive models for sample extremes differ according to (a) whether a Bayesian or non-Bayesian approach is taken, (b) concentration on sample maxima/minima versus threshold exceedences, and (c) the method of fitting and inference. In this article we focus on Bayesian inference for sample maxima/minima where the response is modelled using the Generalized Extreme Value (GEV) distribution. A Bayesian approach offers several advantages, such as automatic estimation of smoothing parameters, incorporation of variability stemming from their estimation, and cogent handling of complications such as missingness. We introduce a new computational method for fitting Bayesian Generalized Extreme Value additive models: Mean Field Variational Bayes. This method has the advantage that Bayesian models can be fitted and assessed quite quickly compared with the established Markov Chain Monte Carlo approach. We expand on this point below.

Figure 1 displays part of a data set that is amenable to GEV additive model analysis. The panels show the time series of annual winter maximum daily rainfall (April–September) at each of 50 weather stations in the hinterland of Sydney, Australia. The data came from a Queensland Department of the Environment (www.derm.qld.gov.au) climate data repository and are known as SILO patched point data. Jeffrey *et al.* (2001) provide details on SILO patched point data. The data set contains several candidate predictor variables, such as geographical location and Southern Oscillatory Index. Preliminary analyses suggest nonlinear

relationships between mean maximum rainfall and these predictor variables. We return to these data, and give the results of a GEV additive model analysis, in Section 6.

Mean Field Variational Bayes (MFVB) (e.g. Attias 1999) is gaining popularity as an alternative to Markov Chain Monte Carlo (MCMC) for Bayesian inference. Recent examples include Braun & McAuliffe (2010), Faes, Ormerod & Wand (2011) and Ren *et al.* (2011). The choice between MFVB and MCMC is not governed by accuracy considerations, since the latter method can always be made more accurate in terms of closeness to the exact posterior distributions. Rather, it is governed by practicality because, depending on the application, MCMC can be unacceptably slow. Contemporary climate research data sets, regarding which GEV additive models have much to offer, tend to be large and complex. The data set analysed in Section 6 contains 1874 records on 8 fields. Fitting just one additive model via the MCMC package BRUGS (Ligges *et al.* 2009) takes several hours on standard hardware platforms. Implementation of MFVB approximations to the same models in the high-level language R (R Development Core Team 2010) reduces fitting times to minutes. Low-level implementation offers even larger speed pay-offs. Such speed is crucial for model fitting and assessment, and opens up the possibility of on-line applications of GEV additive models. MCMC can be contemplated for checking finalized models and improved inference accuracy. However, if WINBUGS is employed for these tasks, the computing times are well over 24 hours on the third author's laptop computer (Max OS X Version 10.6.4, 2.66-GHz processing and 4 GB of random access memory).

Our approach makes use of methodology from Wand *et al.* (2011) for handling GEV distributions within the MFVB framework. Wand *et al.* (2011) explained how the *locality property* of MFVB allows results for the simple univariate models treated there to be incorporated into more complex models involving GEV distributions, such as the GEV additive models looked at here.

Section 2 describes Bayesian GEV additive models. Mean field variational inference for such models is treated in Section 3. In Section 4 we provide guidance on displaying the additive model fits, and in Section 5 we describe extension to geoadditive models. Application to the Sydney hinterland maximum rainfall data is given in Section 6. In Section 7 we compare the new MFVB approach with that based on MCMC. Section 8 contains some closing discussion. All derivations are deferred to an appendix.

1.1. Notation

For each $a \geq 0$, $b \in \mathbb{R}$ and $c > 0$, define the integral

$$\mathcal{J}^+(a, b, c) = \int_0^\infty x^a \exp(bx - cx^2)\, dx. \tag{1}$$

Column vectors with entries consisting of subscripted variables are denoted by a bold-faced version of the letter for that variable. Round brackets will be used to denote the entries of column vectors. For example, $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ denotes an $n \times 1$ vector with entries $x_1, \ldots, x_n$. We use $\mathbf{1}_d$ to denote the $d \times 1$ column vector with all entries equal to 1. The norm of a column vector $\boldsymbol{v}$, defined to be $(\boldsymbol{v}^\top \boldsymbol{v})^{1/2}$, is denoted by $\|\boldsymbol{v}\|$.

The density function of a random vector $\boldsymbol{u}$ is denoted by $p(\boldsymbol{u})$. The conditional density of $\boldsymbol{u}$ given $\boldsymbol{v}$ is denoted by $p(\boldsymbol{u}|\boldsymbol{v})$. The covariance matrix of $\boldsymbol{u}$ is denoted by $\mathrm{cov}(\boldsymbol{u})$.

A random variable x has a GEV distribution with location, scale and shape parameters $\mu, \sigma > 0$ and $\xi \neq 0$, denoted by $x \sim \text{GEV}(\mu, \sigma, \xi)$, if its density function is

$$p(x) = \frac{1}{\sigma} f_{\text{GEV}}\left(\frac{x-\mu}{\sigma}; \xi\right), \tag{2}$$

where

$$f_{\text{GEV}}(x; \xi) = (1 + \xi\, x)^{-1/\xi - 1} \exp\{-(1 + \xi\, x)^{-1/\xi}\}, \quad 1 + \xi\, x > 0 \tag{3}$$

is the GEV$(0, 1, \xi)$ density function. Note that the limiting case $\xi \to 0$ corresponds to the standard Gumbel density function

$$f_{\text{GEV}}(x; 0) = \exp\{-x - \exp(-x)\}, \quad -\infty < x < \infty.$$

A random variable x has a finite normal mixture distribution with location parameter μ, scale parameter $\sigma > 0$ and mixture parameters $\boldsymbol{w} = (w_1, \ldots, w_{\mathcal{K}})$, $\boldsymbol{m} = (m_1, \ldots, m_{\mathcal{K}})$, $\boldsymbol{s} = (s_1, \ldots, s_{\mathcal{K}})$ if its density function is

$$p(x) = \frac{1}{\sigma} f_{\text{NM}}\left(\frac{x-\mu}{\sigma}; \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}\right) \quad \text{where} \quad f_{\text{NM}}(x) = \sum_{k=1}^{\mathcal{K}} \frac{w_k}{s_k} \phi\left(\frac{x - m_k}{s_k}\right)$$

and $\phi(z) = (2\pi)^{-1/2} e^{-z^2/2}$. The $\boldsymbol{w}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ vectors are, respectively, the weights, means and standard deviations of the mixture components. Note the restrictions $\sum_{k=1}^{\mathcal{K}} w_k = 1$ and the $w_k, s_k > 0$ for each $1 \leq k \leq \mathcal{K}$. Throughout this article we take $\boldsymbol{w}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ to be fixed known vectors, whereas μ and σ are treated as unknown parameters. We write

$$x \sim \text{Normal-Mixture}(\mu, \sigma; \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}).$$

A random variable x has an inverse gamma distribution with parameters $A, B > 0$, denoted by $x \sim \text{IG}(A, B)$ if its density function is $p(x) = B^A \Gamma(A)^{-1} x^{-A-1} e^{-B/x}$, $x > 0$. The $\mathcal{K} \times 1$ random vector $\boldsymbol{x}$ has a Multinomial$(1; \boldsymbol{\pi})$ distribution, where $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_{\mathcal{K}})$ is such that $\sum_{k=1}^{\mathcal{K}} \pi_k = 1$, if its probability mass function is $p(x_1, \ldots, x_K) = \prod_{k=1}^{K} \pi_k^{x_k}$, $x_k = 0, 1$, for $1 \leq k \leq \mathcal{K}$.

If y_i has distribution D_i for each $1 \leq i \leq n$, and the y_i are independent, then we write $y_i \overset{\text{ind.}}{\sim} D_i$.

2. Bayesian Generalized Extreme Value additive models

Let y_i, $1 \leq i \leq n$, be a set of response variables for which a GEV$(\mu_i, \sigma_\varepsilon, \xi)$ distribution is appropriate. Typically the y_i represent either maxima or minima for each of n samples. GEV additive models assume that the locations take the form

$$\mu_i = f_1(x_{1i}) + \cdots + f_d(x_{di}), \tag{4}$$

where, for each $1 \leq i \leq n$, $(x_{1i}, \ldots, x_{di})$ is a vector of continuous predictor variables and the $f_1, \ldots, f_d$ are smooth, but otherwise arbitrary, functions.

There are numerous ways by which the f_ℓ in (4) can be modelled and estimated. The mixed model-based penalized spline approach (e.g. Padoan & Wand 2008) has a natural Bayesian representation, and thus allows Bayesian inference tools such as Markov Chain

Monte Carlo and MFVB to be used for fitting and inference. This involves modelling the right-hand side of (4) as

$$\sum_{\ell=1}^{d} f_\ell(x_\ell) = \beta_0 + \sum_{\ell=1}^{d} \left\{ \beta_\ell x_\ell + \sum_{k=1}^{K_\ell} u_{\ell,k} z_{\ell,k}(x_\ell) \right\}$$

with

$$u_{\ell,1}, \ldots, u_{\ell,K_\ell} | \sigma_\ell^2 \overset{\text{ind.}}{\sim} \text{N}\big(0, \sigma_\ell^2\big) \quad \text{for each } 1 \le \ell \le d.$$

Here $\{z_{\ell,1}(\cdot), \ldots, z_{\ell,K_\ell}(\cdot)\}$ is a set of spline basis functions for estimation of f_ℓ. See, for example, Welham *et al.* (2007) and Wand & Ormerod (2008) for the construction of such functions. Define the matrices

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}, \qquad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix},$$

$$\boldsymbol{u}_\ell = \begin{bmatrix} u_{\ell,1} \\ \vdots \\ u_{\ell,K_\ell} \end{bmatrix}, \qquad \boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_d \end{bmatrix}, \qquad \boldsymbol{Z}_\ell = \begin{bmatrix} z_{\ell,1}(x_{1\ell}) & \cdots & z_{\ell,K_\ell}(x_{1\ell}) \\ \vdots & \ddots & \vdots \\ z_{\ell,1}(x_{n\ell}) & \cdots & z_{\ell,K_\ell}(x_{n\ell}) \end{bmatrix}$$

and then set $\boldsymbol{Z} = [\boldsymbol{Z}_1 \cdots \boldsymbol{Z}_d]$. Then a Bayesian GEV additive model is

$$y_i | \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon, \xi \overset{\text{ind.}}{\sim} \text{GEV}((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i, \sigma_\varepsilon, \xi)$$

$$\boldsymbol{u} \big| \sigma_{u1}^2, \ldots, \sigma_{ud}^2 \sim \text{N}\big(\boldsymbol{0}, \text{blockdiag}\big(\sigma_{u1}^2 \boldsymbol{I}, \ldots, \sigma_{ud}^2 \boldsymbol{I}\big)\big) \tag{5}$$

$$\boldsymbol{\beta} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_\beta), \quad \sigma_\varepsilon^2 \sim \text{IG}(A_\varepsilon, B_\varepsilon), \quad \sigma_{u\ell}^2 \overset{\text{ind.}}{\sim} \text{IG}(A_{u\ell}, B_{u\ell}), \quad \xi \sim p(\xi),\ \xi \in \Xi,$$

where $\boldsymbol{\Sigma}_\beta$ is a symmetric and positive-definite $(d+1) \times (d+1)$ matrix, and $A_\varepsilon, B_\varepsilon, A_{u\ell}, B_{u\ell} > 0$ are hyperparameters for the variance component prior distributions. We impose a discrete finite prior distribution on ξ. The set over which the prior probability mass function $p(\xi)$ is positive is denoted by Ξ. Figure 2 shows Bayesian model (5), for $d = 3$, as a directed acyclic graph. Arrows convey conditional dependence between the components of the model. As illustrated by (17) in the Appendix, MFVB calculations benefit from such graphical representation of Bayesian model (5).

The last line of (5) lists prior distributions for the model parameters. Their forms are chosen to enhance the tractability of MFVB fitting and inference. Other families of priors may be used, but lead to algorithms requiring non-analytic integrals (Wand *et al.* 2011). The imposition of a discrete finite prior on ξ means that MFVB inference can be done for each

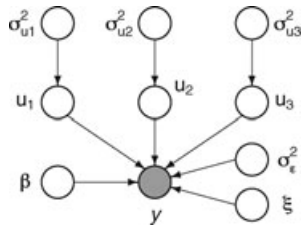

Figure 2. Directed acylic graph representation of the Bayesian generalized extreme value additive model (5) in the case of $d = 3$.

value of $\xi \in \Xi$ and the results combined at the end. This is an effective means of handling this parameter, as its appearance in the likelihood involves complicated forms that are not amenable to direct MFVB calculations.

Before describing the MFVB fitting of Bayesian model (5) we note that it can be fit straightforwardly via the MCMC software package WINBUGS, although we are not aware of any published work that has described such implementation. The WINBUGS code is similar to that used in the examples in Crainiceanu, Ruppert & Wand (2005), Gurrin, Scurrah & Hazelton (2005) and Marley & Wand (2010). The likelihood specification $y_i \sim \text{GEV}(\mu_i, \sigma_\varepsilon, \xi)$ takes the form

$$\texttt{y[i]} \sim \texttt{dGEV(mu[i], sigmaEpsilon, xi)}$$

where `dGEV` corresponds to the form of the GEV family of density functions given by (2) and (3).

Before closing this section, we reiterate that maxima and minima are not the only response variables used in regression analysis for extreme events. An alternative approach involves taking the response variable to be the number in the sample for which a certain threshold is exceeded and is often referred to as a *peaks-over-threshold* analysis. The generalized Pareto distribution typically is used to model such responses. Chavez-Demoulin & Davison (2005) developed generalized additive models of this type.

3. Mean Field Variational Bayes inference

Our approach to MFVB inference for GEV additive models consists of three stages. The first stage involves a finite normal mixture approximation of the GEV$(0, 1, \xi)$ density function over each $\xi \in \Xi$. The second stage involves MFVB inference for finite normal mixture additive models. Such models take the same form as model (5), but with a finite normal mixture distribution used to model the responses. Section 3.2 describes such models and a MFVB fitting algorithm. For fixed ξ this algorithm results in approximate posteriors for the additive model parameters $\boldsymbol{\beta}$, $\boldsymbol{u}$, σ_ε and $\sigma^2_{u\ell}$, $1 \le \ell \le d$. The final stage is to combine the results across all fits to make approximate Bayesian inference for all model parameters, including the shape parameter ξ.

This *auxiliary mixture* approach to handling GEV responses was developed by Wand *et al.* (2011), although only simple univariate GEV models were treated there. Appendix C of Wand *et al.* (2011) contains details on normal mixture approximation of the $f_{\text{GEV}}(\cdot; \xi)$ and

provides access to the $\mathcal{K} = 24$ solutions for the shape parameter values

$$\xi \in \{-1, -0.995, \ldots, 0.995, 1\}.$$

Therefore, provided we take Ξ to be a subset of these values, we can make use of the same normal mixture approximations for MFVB fitting of GEV additive models. In applications involving sample extrema, the restriction $-1 \leq \xi \leq 1$ is usually adequate. In the application of Section 6 we found $\Xi = \{0, 0.01, \ldots, 0.49, 0.5\}$ to be adequate. This is consistent with Koutsoyiannis (2004), who obtained estimated ξ values in the vicinity of 0.1 for maximum rainfall data.

3.1. Basic principles of Mean Field Variational Bayes

Here we describe the basic principles of MFVB in brief. Fuller details can be found in Bishop (2006, chapter 10) and Ormerod & Wand (2010).

Consider a generic Bayesian model, with observed data vector $\boldsymbol{y}$ and parameter vector $\boldsymbol{\theta}$. Suppose that $\boldsymbol{\theta}$ is continuous over the parameter space $\boldsymbol{\Theta}$. The treatment for discrete parameter spaces is similar. The posterior density function $p(\boldsymbol{\theta}|\boldsymbol{y})$ is often intractable. MFVB overcomes intractability by postulating that $p(\boldsymbol{\theta}|\boldsymbol{y})$ can be well approximated by product density forms. An example is

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \approx q_1(\boldsymbol{\theta}_1)\, q_2(\boldsymbol{\theta}_2)\, q_3(\boldsymbol{\theta}_3), \tag{6}$$

where $\{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$ is a partition of the entries of $\boldsymbol{\theta}$. Obviously, there are numerous ways by which $\boldsymbol{\theta}$ can be partitioned, and the choice of partition is usually made on tractability grounds. Each q_i is a density function in $\boldsymbol{\theta}_i$ ($i = 1, 2, 3$), and they are chosen to minimize the Kullback–Leibler distance between the left- and right-hand sides of (6):

$$\int q_1(\boldsymbol{\theta}_1)\, q_2(\boldsymbol{\theta}_2)\, q_3(\boldsymbol{\theta}_3) \log \left\{ \frac{p(\boldsymbol{\theta}|\boldsymbol{y})}{q_1(\boldsymbol{\theta}_1)\, q_2(\boldsymbol{\theta}_2)\, q_3(\boldsymbol{\theta}_3)} \right\} d\boldsymbol{\theta}. \tag{7}$$

Minimization of (7) is equivalent to maximization of

$$\underline{p}(\boldsymbol{y}; q) = \int q_1(\boldsymbol{\theta}_1)\, q_2(\boldsymbol{\theta}_2)\, q_3(\boldsymbol{\theta}_3) \log \left\{ \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{q_1(\boldsymbol{\theta}_1)\, q_2(\boldsymbol{\theta}_2)\, q_3(\boldsymbol{\theta}_3)} \right\} d\boldsymbol{\theta}$$

(e.g. Bishop 2006, section 10.1) and an iterative convex optimization algorithm (e.g. Luenberger & Ye 2008) is available for obtaining the solution. The algorithm updates can be derived from the following conditions on the solutions, q_1^*, q_2^* and q_3^*:

$$\begin{aligned} q_1^*(\theta_1) &\propto \exp E_{q_2(\theta_2)q_3(\theta_3)}\{\log\ p(\theta_1|\boldsymbol{y}, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)\}, \\ q_2^*(\theta_2) &\propto \exp E_{q_1(\theta_2)q_3(\theta_3)}\{\log\ p(\theta_2|\boldsymbol{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_3)\}, \\ q_3^*(\theta_3) &\propto \exp E_{q_1(\theta_2)q_2(\theta_3)}\{\log\ p(\theta_3|\boldsymbol{y}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)\}. \end{aligned}$$

Each iteration results in an increase in $\underline{p}(\boldsymbol{y}; q)$, and this quantity can be used for assessment of convergence. Moreover, upon convergence, $\underline{p}(\boldsymbol{y}; q)$ approximates the marginal likelihood $p(\boldsymbol{y})$. In practice, it is prudent to work with $\log\ \underline{p}(\boldsymbol{y}; q)$ to avoid underflow/overflow.

Upon convergence, the q_i^* densities can be used for approximate Bayesian inference. The quality of the approximation depends on the reasonableness of (6). The treatment of

other partitions of $\boldsymbol{\theta}$ is analogous. The partition is usually chosen to enhance tractability while keeping product density restrictions to a minimum.

Typically we drop the asterisks and the subscripts on the q_i^* and let the argument signify the relevant parameter. Finally, the notation

$$\boldsymbol{\mu}_{q(\boldsymbol{\theta})} = E_q(\boldsymbol{\theta}) \quad \text{and} \quad \boldsymbol{\Sigma}_{q(\boldsymbol{\theta})} = \text{cov}_q(\boldsymbol{\theta})$$

is quite useful for describing the updates. In the scalar case we use

$$\mu_{q(\theta)} = E_q(\theta) \quad \text{and} \quad \sigma^2_{q(\theta)} = \text{var}_q(\boldsymbol{\theta}).$$

3.1.1. *Extension to Structured Mean Field Variational Bayes*

Our approach to GEV additive model analysis requires an extension of the MFVB paradigm known as *structured* MFVB (Saul & Jordan 1996). This involves restriction of the GEV shape parameter, ξ, to a finite set Ξ and applying MFVB conditionally for $\xi \in \Xi$. Section 3.1 of Wand *et al.* (2011) contains details on structured MFVB approximation.

3.2. Finite normal mixture responses

Consider the model

$$\begin{aligned} y_i|\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon, \xi &\overset{\text{ind.}}{\sim} \text{Normal-Mixture}\left((\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u})_i, \sigma_\varepsilon, \boldsymbol{w}, \boldsymbol{m}, \boldsymbol{s}\right), \\ \boldsymbol{u}|\sigma^2_{u1}, \ldots, \sigma^2_{ud} &\sim \text{N}\big(\boldsymbol{0}, \text{blockdiag}\big(\sigma^2_{u1}\boldsymbol{I}, \ldots, \sigma^2_{ud}\boldsymbol{I}\big)\big), \\ \boldsymbol{\beta} \sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\ \sigma^2_\varepsilon &\sim \text{IG}(A_\varepsilon, B_\varepsilon),\ \sigma^2_{u\ell} \overset{\text{ind.}}{\sim} \text{IG}(A_{u\ell}, B_{u\ell}). \end{aligned} \tag{8}$$

In model (8), $\boldsymbol{w}$, $\boldsymbol{m}$ and $\boldsymbol{s}$ are each *fixed* vectors and do not require Bayesian inference. Hence, we are *not* concerned with the classical normal mixture fitting problem in this section. Model (8) is not of interest in its own right but, rather, is a stepping stone towards MFVB inference for GEV additive models. As explained in Wand *et al.* (2011), we can re-express (8) as

$$\begin{aligned} p(y_i|a_i, \boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon) &= \prod_{k=1}^{\mathcal{K}} \left[\frac{1}{\sigma_\varepsilon s_k}\,\phi\left(\frac{\sigma_\varepsilon^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i - m_k}{s_k}\right)\right]^{a_{ik}}, \\ \boldsymbol{a}_i &\overset{\text{ind.}}{\sim} \text{Multinomial}(1, \boldsymbol{w}), \quad \boldsymbol{u}\big|\sigma^2_{u1}, \ldots, \sigma^2_{ud} \sim \text{N}\big(\boldsymbol{0}, \text{blockdiag}\big(\sigma^2_{u1}\boldsymbol{I}, \ldots, \sigma^2_{ud}\boldsymbol{I}\big)\big), \\ \boldsymbol{\beta} &\sim \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}),\ \sigma^2_\varepsilon \sim \text{IG}(A_\varepsilon, B_\varepsilon),\ \sigma^2_{u\ell} \overset{\text{ind.}}{\sim} \text{IG}(A_{u\ell}, B_{u\ell}), \end{aligned} \tag{9}$$

where $\boldsymbol{a}_i$, $1 \le i \le n$, is a set of $\mathcal{K} \times 1$ auxiliary variables.

Consider MFVB fitting of (9) with the product restriction

$$q\big(\boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_\varepsilon, \sigma^2_{u1}, \ldots, \sigma^2_{ud}, \boldsymbol{a}\big) = q(\boldsymbol{\beta}, \boldsymbol{u})\, q\big(\sigma^2_\varepsilon, \sigma^2_{u1}, \ldots, \sigma^2_{ud}\big) q(\boldsymbol{a}).$$

Then the optimal densities take the form

$$\begin{aligned} q^*(\boldsymbol{\beta}, \boldsymbol{u}) &\sim \mathrm{N}(\mu_{q(\boldsymbol{\beta},\boldsymbol{u})}, \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}), \\ q^*\big(\sigma^2_{u\ell}\big) &\sim \mathrm{IG}\left(A_{u\ell} + \frac{1}{2}K_\ell, B_{q(\sigma^2_{u\ell})}\right), \\ q^*\big(\sigma^2_\varepsilon\big) &= \frac{\big(\sigma^2_\varepsilon\big)^{-A-\frac{1}{2}n-1}\exp\big\{C_9/\big(\sigma^2_\varepsilon\big)^{1/2} - C_{10}/\sigma^2_\varepsilon\big\}}{2\mathcal{J}^+(2A+n-1, C_9, C_{10})}, \quad \sigma^2_\varepsilon > 0, \\ \text{and} \quad q^*(\boldsymbol{a}_i) &\overset{\text{ind.}}{\sim} \text{Multinomial}(1; \boldsymbol{\mu}_{q(\boldsymbol{a}_i)}), \end{aligned} \tag{10}$$

where the parameters in (10) can be obtained iteratively via Algorithm 3.2. The derivation of the updates is given in an appendix. The algorithm uses the following notation:

$$\boldsymbol{C} = [\boldsymbol{X}\ \boldsymbol{Z}], \quad \boldsymbol{D}_{\boldsymbol{a}_k} = \operatorname*{diag}_{1\le i\le n}(a_{ik}), \quad \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}} = \operatorname*{diag}_{1\le i\le n}\{\mu_{q(a_{ik})}\}.$$

Algorithm 3.2 Iterative scheme for obtaining the parameters in the optimal densities $q^*(a)$, $q^*(\beta, u)$, $q^*(\sigma^2_\varepsilon)$ and $q^*(\sigma^2_{u\ell})$ for the finite normal mixture model.

Initialize: $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}$, $\mu_{q(1/\sigma_\varepsilon)}$, $\mu_{q(1/\sigma^2_\varepsilon)}$

Cycle:

Update $q^*(\boldsymbol{a})$ parameters:

For $i = 1, \ldots, n$, $k = 1, \ldots, \mathcal{K}$:

$$\begin{aligned} \nu_{ik} \leftarrow \log(w_k/s_k) - \frac{1}{2s^2_k}\Big[\mu_{q(1/\sigma^2_\varepsilon)}\big\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})^2_i + (\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top)_{ii}\big\} \\ - 2\mu_{q(1/\sigma_\varepsilon)}m_k(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})_i + m^2_k\Big]. \end{aligned}$$

For $i = 1, \ldots, n$, $k = 1, \ldots, \mathcal{K}$: $\mu_{q(a_{ik})} \leftarrow \exp(\nu_{ik}) \Big/ \sum_{k=1}^{\mathcal{K}} \exp(\nu_{ik})$

For $k = 1, \ldots, \mathcal{K}$: $\mu_{q(a_{\bullet k})} \leftarrow \sum_{i=1}^{n} \mu_{q(a_{ik})}$

Update $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ parameters:

$$\begin{aligned} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \leftarrow \Bigg\{ \boldsymbol{C}^\top \left(\mu_{q(1/\sigma^2_\varepsilon)} \sum_{k=1}^{\mathcal{K}} \frac{1}{s^2_k} \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\right) \boldsymbol{C} \\ + \text{blockdiag}\big(\boldsymbol{\Sigma}^{-1}_{\boldsymbol{\beta}}, \mu_{q(1/\sigma^2_{u1})}\boldsymbol{I}_{K_1}, \ldots, \mu_{q(1/\sigma^2_{ud})}\boldsymbol{I}_{K_d}\big)\Bigg\}^{-1} \end{aligned}$$

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{\mu})} \leftarrow \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\left[\boldsymbol{C}^\top\left\{\mu_{q(1/\sigma^2_\varepsilon)}\sum_{k=1}^{\mathcal{K}}\frac{1}{s^2_k}\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{y} - \mu_{q(1/\sigma_\varepsilon)}\sum_{k=1}^{\mathcal{K}}\frac{m_k}{s^2_k}\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{1}\right\}\right]$$

Update $q^*(\sigma_\varepsilon^2)$ parameters:

$$C_9 \leftarrow \sum_{k=1}^{\mathcal{K}} \frac{m_k}{s_k^2} \mathbf{1}^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})$$

$$C_{10} \leftarrow B_\varepsilon + \frac{1}{2} \sum_{k=1}^{\mathcal{K}} \frac{1}{s_k^2} \Big\{ \operatorname{tr}\left(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})} \boldsymbol{C}^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}} \boldsymbol{C}\right) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}} (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})}) \Big\}$$

$$\mu_{q(1/\sigma_\varepsilon^2)} \leftarrow \frac{\mathcal{J}^+(2A_\varepsilon + n + 1, C_9, C_{10})}{\mathcal{J}^+(2A_\varepsilon + n - 1, C_9, C_{10})}; \quad \mu_{q(1/\sigma_\varepsilon)} \leftarrow \frac{\mathcal{J}^+(2A_\varepsilon + n, C_9, C_{10})}{\mathcal{J}^+(2A_\varepsilon + n - 1, C_9, C_{10})}$$

Update $q^*(\sigma_{u\ell}^2)$ parameters:

For $\ell = 1, \ldots, d$:

$$B_{q(\sigma_{u\ell}^2)} \leftarrow B_{u\ell} + \frac{1}{2} \left\{ \|\boldsymbol{\mu}_{q(\mu_\ell)}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)}) \right\}; \quad \mu_{q(1/\sigma_{u\ell}^2)} \leftarrow \left(A_{u\ell} + \frac{1}{2} K_\ell \right) \Big/ B_{q(\sigma_{u\ell}^2)}$$

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

The C_9 and C_{10} notation matches that used by Wand *et al.* (2011) for the univariate normal mixture model. An expression for $\log \underline{p}(\boldsymbol{y}; q)$, valid after all of the updates in Algorithm 3.2 have been made, is

$$\begin{aligned}
\log \underline{p}(\boldsymbol{y}; q) = {} & \frac{1}{2} \left(1 + d + \sum_{\ell=1}^{d} K_\ell \right) - \frac{n}{2} \log(2\pi) + \log(2) + A_\varepsilon \log(B_\varepsilon) - \log \Gamma(A_\varepsilon) \\
& + \log \mathcal{J}^+(2A_\varepsilon + n - 1, C_9, C_{10}) \\
& + \frac{1}{2} \log |\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\mu})}| - \frac{1}{2} \log |\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| - \frac{1}{2} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}{}^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \frac{1}{2} \operatorname{tr}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta})} \right) \\
& + \sum_{\ell=1}^{d} \left\{ A_{u\ell} \log(B_{u\ell}) - \left(A_{u\ell} + \frac{K_\ell}{2} \right) \log\left(B_{q(\sigma_{u\ell}^2)} \right) + \log \Gamma\left(A_{u\ell} + \frac{K_\ell}{2} \right) \right. \\
& \left. - \log \Gamma(A_{u\ell}) \right\} + \sum_{k=1}^{\mathcal{K}} \mu_{q(a_{\bullet k})} \left\{ \log(w_k / s_k) - \frac{1}{2} \left(m_k^2 / s_k^2 \right) \right\} \\
& - \sum_{i=1}^{n} \sum_{k=1}^{\mathcal{K}} \mu_{q(a_{ik})} \log(\mu_{q(a_{ik})}).
\end{aligned} \tag{11}$$

Finally, note that $\mathcal{J}^+(a, b, c)$, defined by (1), does not admit a closed-form solution for general $a \geq 0$. Appendix B of Wand *et al.* (2011) provides an efficient quadrature scheme for its evaluation. The scheme involves Laplace approximation of the integrand to obtain its closest normal density function. This is followed by centring and scaling of the integral so that its support approximately matches that of the N(0, 1) density function, which facilitates selection of a good initial quadrature grid. Adaptive trapezoidal quadrature is then employed. An alternative route for computation of $\mathcal{J}^+(a, b, c)$ is to note that it can be expressed in terms

of a *parabolic cylinder function*. Specifically,

$$\mathcal{J}^{+}(a,b,c) = (2r)^{-(a+1)/2}\Gamma(a+1)\exp\{b^2/(8c)\}D_{-a-1}\left(-\frac{b}{(2c)^{1/2}}\right), \tag{12}$$

where D_ν is the parabolic cylinder function of order ν as defined in Gradshteyn & Ryzhik (1994). However, care needs to be taken with (12) to avoid underflow/overflow.

3.3. Generalized Extreme Value responses

Recall that $f_{\mathrm{GEV}}(\cdot;\xi)$ denotes the GEV(0, 1, ξ) family of density functions, and is given by (3). As described in Wand *et al.* (2011), we can replace each $f_{\mathrm{GEV}}(\cdot;\xi)$, $\xi \in \Xi$, by a highly accurate normal mixture approximation:

$$f_{\mathrm{GEV}}(x;\xi) \approx \sum_{k=1}^{\mathcal{K}} \frac{w_{k,\xi}}{s_{k,\xi}}\,\phi\left(\frac{x-m_{k,\xi}}{s_{k,\xi}}\right). \tag{13}$$

Given approximation (13), for each fixed $\xi \in \Xi$ we can use Algorithm 3.2 to obtain MFVB approximations, with the restrictions

$$q\big(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{u1}^2, \ldots, \sigma_{ud}^2, \xi\big) = q(\xi)q(\boldsymbol{\beta}, \boldsymbol{u}|\xi)q\big(\sigma_\varepsilon^2, \sigma_{u1}^2, \ldots, \sigma_{ud}^2\big|\xi\big).$$

Let these approximations be denoted by

$$q^*(\boldsymbol{\beta}, \boldsymbol{u}|\xi) \quad \text{and} \quad q^*\big(\sigma_\varepsilon^2, \sigma_{u1}^2, \ldots, \sigma_{ud}^2\big|\xi\big) = q^*\big(\sigma_\varepsilon^2\big|\xi\big)q^*\big(\sigma_{u1}^2\big|\xi\big)\cdots q^*\big(\sigma_{ud}^2\big|\xi\big)$$

respectively.

Using results from section 3.1 of Wand *et al.* (2011), with $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{u1}^2, \ldots, \sigma_{ud}^2)$ and $\eta = \xi$, we obtain the *structured* MFVB approximations

$$q^*(\xi) = \frac{p(\xi)\,\underline{p}(\boldsymbol{y}|\xi)}{\displaystyle\sum_{\xi'\in\Xi} p(\xi')\,\underline{p}(\boldsymbol{y}|\xi')}, \quad q^*(\boldsymbol{\beta}, \boldsymbol{u}) = \sum_{\xi\in\Xi} q^*(\xi)\,q^*(\boldsymbol{\beta}, \boldsymbol{u}\big|\xi),$$

$$q^*(\sigma_\varepsilon^2) = \sum_{\xi\in\Xi} q^*(\xi)\,q^*(\sigma_\varepsilon^2\big|\xi)$$

and

$$p\big(\sigma_{u\ell}^2\big|\boldsymbol{y}\big) = \sum_{\xi\in\Xi} q^*(\xi)p\big(\sigma_{u\ell}^2\big|\boldsymbol{y}, \xi\big), \qquad 1 \le \ell \le d.$$

The approximate marginal log-likelihood is

$$\underline{p}(\boldsymbol{y}; q) = \sum_{\xi\in\Xi} q^*(\xi)\underline{p}(\boldsymbol{y}|\xi).$$

In view of (10) it is apparent that $q^*(\boldsymbol{\beta}, \boldsymbol{u})$ is a finite multivariate normal mixture density function with weights given by the approximate posterior probability mass function of ξ. Similarly, $q^*(\sigma_{u\ell}^2)$ is a finite mixture of inverse gamma density functions with the same set of weights. The form for $q^*(\sigma_\varepsilon^2)$ is analogous, but involves the non-standard form apparent in (10).

Algorithm 3.3 summarizes this finite normal mixture approach to structured MFVB inference for $(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_\varepsilon^2, \sigma_{u1}^2, \ldots, \sigma_{ud}^2, \xi)$ in the GEV additive model (5). The algorithm assumes that finite normal mixture approximations of the form (13) have been obtained for each $\xi \in \Xi$.

Algorithm 3.3 Scheme for structured MFVB approximation of the posterior distributions of parameters in the GEV additive model.
For each $\xi \in \Xi$:

1. Retrieve the normal mixture approximation vectors: $(w_{k,\xi}, m_{k,\xi}, s_{k,\xi})$, $1 \le k \le \mathcal{K}$, for approximation of the GEV(0, 1, ξ) density function.
2. Apply Algorithm 3.2 with (w_k, m_k, s_k) set to $(w_{k,\xi}, m_{k,\xi}, s_{k,\xi})$, $1 \le k \le \mathcal{K}$.
3. Store the parameters needed to define $q^*(\boldsymbol{\beta}, \boldsymbol{u}|\xi)$, $q^*(\sigma_\varepsilon^2|\xi)$ and $q^*(\sigma_{u\ell}^2|\xi)$, $1 \le \ell \le d$.
4. Store the converged marginal likelihood lower bound $\underline{p}(\boldsymbol{y}|\xi)$.

Form the approximations to the posteriors $p(\xi|\boldsymbol{y})$, $p(\boldsymbol{\beta}, \boldsymbol{u}|\boldsymbol{y})$, $p(\sigma_\varepsilon^2|\boldsymbol{y})$ and $p(\sigma_{u\ell}^2|\boldsymbol{y})$:

$$q^*(\xi) = \frac{p(\xi)\,\underline{p}(\boldsymbol{y}|\xi)}{\sum_{\xi' \in \Xi} p(\xi')\,\underline{p}(\boldsymbol{y}|\xi')}, \quad q^*(\boldsymbol{\beta}, \boldsymbol{u}) = \sum_{\xi \in \Xi} q^*(\xi)\, q^*(\boldsymbol{\beta}, \boldsymbol{u}|\xi),$$

$$q^*\big(\sigma_\varepsilon^2\big) = \sum_{\xi \in \Xi} q^*(\xi)\, q^*\big(\sigma_\varepsilon^2\big|\xi\big) \quad \text{and} \quad q^*\big(\sigma_{u\ell}^2\big) = \sum_{\xi \in \Xi} q^*(\xi)\, q^*\big(\sigma_{u\ell}^2\big|\xi\big), \quad 1 \le \ell \le d.$$

Form the approximate marginal likelihood: $\underline{p}(\boldsymbol{y}; q) = \sum_{\xi \in \Xi} q^*(\xi)\underline{p}(\boldsymbol{y}|\xi)$.

---

4. Displaying additive model fits

The previous section describes MFVB methodology for obtaining the approximate posterior distributions of the model parameters. However, conversion of these into meaningful graphical displays requires additional non-trivial manipulations, which we now describe.

Consider the problem of displaying the first fitted function, $\widehat{f_1}$, over a grid of M points $\boldsymbol{g}_1 = (g_{11}, \ldots, g_{1M})$. The g_{1j} can be quite general, although typically they are equi-spaced and encompass the x_{1i}, $1 \le i \le n$, values. To align the vertical axis with the response data it is recommended that grids for the other variables be set to $\boldsymbol{g}_\ell = \bar{x}_\ell \mathbf{1}_M$ for $2 \le \ell \le d$, where $\mathbf{1}_M$ is the $M \times 1$ vector of ones. This choice leads to the display for the first fitted curve being a slice of the d-variate fitted function with all other dimensions set at the average predictor values. Set

$$\boldsymbol{X}_{\boldsymbol{g}}^{(1)} = [\mathbf{1}_M\ \boldsymbol{g}_1\ , \ldots,\ \boldsymbol{g}_d], \qquad \boldsymbol{Z}_{\ell\boldsymbol{g}}^{(1)} = [z_{\ell,1}(\boldsymbol{g}_\ell), \ldots, z_{\ell,K_\ell}(\boldsymbol{g}_\ell)], \quad 1 \le \ell \le d,$$

and then put $\boldsymbol{C}_{\boldsymbol{g}}^{(1)} = [\boldsymbol{X}_{\boldsymbol{g}}^{(1)}|\boldsymbol{Z}_{1\boldsymbol{g}}^{(1)}, \ldots, \boldsymbol{Z}_{d\boldsymbol{g}}^{(1)}]$. Then

$$\boldsymbol{f}_1 = \text{approximate posterior mean of } f_1(\boldsymbol{g}_1) = \boldsymbol{C}_{\boldsymbol{g}}^{(1)} \boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u})} = \sum_{\xi \in \Xi} q^*(\xi) \boldsymbol{C}_{\boldsymbol{g}}^{(1)} \boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u}|\xi)}.$$

A plot of $\boldsymbol{f}_1$ versus $\boldsymbol{g}_1$ provides a display of the fit for f_1, with appropriate vertical alignment.

However, it is also useful to plot posterior pointwise credible intervals. Because 95% credible sets are the most common, the remainder of this section will be devoted to their construction. Credible intervals with other levels require simple adjustment. Pointwise 95% credible intervals, with equal-sized tails, require the 0.025 and 0.975 quantiles of the MFVB approximation of the entries of the $M \times 1$ vector

$$\boldsymbol{C}_g^{(1)} \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}.$$

Since the MFVB approximate posterior density of $(\boldsymbol{\beta}, \boldsymbol{u})$ is a finite normal mixture, we require the following result to obtain these quantiles.

Result 1. *Suppose that the $r \times 1$ random vector $\boldsymbol{x}$ has the finite normal mixture density function*

$$p(\boldsymbol{x}) = \sum_{\ell=1}^{L} \omega_\ell (2\pi)^{-r/2} |\boldsymbol{\Sigma}_\ell|^{-1/2} \exp\left\{ -\frac{1}{2} (\boldsymbol{x} - \boldsymbol{\mu}_\ell)^\top \boldsymbol{\Sigma}_\ell^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_\ell) \right\},$$

where $\sum_{\ell=1}^{L} \omega_\ell = 1$ and, for $1 \le \ell \le L$, $\omega_\ell > 0$, the $\boldsymbol{\mu}_\ell$ are unrestricted $r \times 1$ vectors and the $\boldsymbol{\Sigma}_\ell$ are $r \times r$ symmetric positive-definite matrices. Write this as

$$\boldsymbol{x} \sim \omega_1 \, \mathrm{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) + \cdots + \omega_L \, \mathrm{N}(\boldsymbol{\mu}_L, \boldsymbol{\Sigma}_L).$$

Then, for any constant $r \times 1$ vector $\boldsymbol{\alpha}$,

$$\boldsymbol{\alpha}^\top \boldsymbol{x} \sim \omega_1 \, \mathrm{N}\big(\boldsymbol{\alpha}^\top \boldsymbol{\mu}_1, \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_1 \boldsymbol{\alpha}\big) + \cdots + \omega_L \, \mathrm{N}\big(\boldsymbol{\alpha}^\top \boldsymbol{\mu}_L, \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}_L \boldsymbol{\alpha}\big).$$

For each $1 \le j \le M$ let $\boldsymbol{e}_j$ denote the $M \times 1$ vector having jth entry equal to one and zeroes elsewhere. Then, as a consequence of Result 1, the 95% credible interval limits for the jth grid point are the 0.025 and 0.975 quantiles of the

$$\sum_{\xi \in \Xi} q^*(\xi) \mathrm{N}\left(\boldsymbol{e}_j{}^\top \boldsymbol{C}_g^{(1)} \boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u}|\xi)}, \boldsymbol{e}_j{}^\top \boldsymbol{C}_g^{(1)} \boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u}|\xi)} \big(\boldsymbol{C}_g^{(1)}\big)^\top \boldsymbol{e}_j \right)$$

density function. Therefore, computation of approximate pointwise credible sets requires a routine for obtaining quantiles of general univariate finite normal mixture distributions. Displays for the other fitted functions can be obtained using analogous manipulations.

5. Geoadditive extension

A common extension of the additive model (4) for data sets possessing geographical information is

$$\mu_i = f_1(x_{1i}) + \cdots + f_d(x_{di}) + g(\texttt{longitude}_i, \texttt{latitude}_i), \tag{14}$$

where the additional $g(\texttt{longitude}_i, \texttt{latitude}_i)$ term represents a bivariate function of geographical position. Of course, longitude/latitude can be replaced by any other coordinate system for specification of geographical position. Extension (14), often referred to as *geoadditive models*, has a large and rapidly growing literature. Early contributions include Wahba

et al. (1995), Kammann & Wand (2003), Wood (2003) and Hennerfeind, Brezger & Fahrmeir (2006). One can handle the bivariate function via the mechanism

$$g(\texttt{longitude}, \texttt{latitude}) = \beta_{\text{lon}}\texttt{longitude} + \beta_{\text{lat}}\texttt{latitude} + \sum_{k=1}^{K_g} u_{gk} z_{gk}(\texttt{longitude}, \texttt{latitude})$$

for a suitable set of bivariate basis functions $\{z_{gk} : 1 \leq k \leq K_g\}$ and with their coefficients obeying $u_{gk}|\sigma^2_{ug} \overset{\text{ind.}}{\sim} \text{N}(0, \sigma^2_{ug})$, $1 \leq k \leq K_g$. It follows that the GEV geoadditive models can also be fitted via Algorithms 3.2 and 3.3, with an additional variance component for geographical location.

6. Application

We now provide illustration of the methodology via the Sydney hinterland maximum rainfall data. The following variables were considered:

`winter max. rainfall:`	maximum rainfall (mm) for annual winter period, defined as April to September inclusive;
`year:`	year (1955–2003);
`day in season:`	day of winter period (i.e. number of days since 31st March within the current year);
`OHA:`	Ocean Heat content Anomaly (10^{22} joules);
`SOI:`	Southern Oscillatory Index;
`PDO:`	Pacific Decadal Oscillation;
`longitude:`	degrees longitude of weather station;
`latitude:`	degrees latitude of weather station.

The predictor variables OHA, SOI and PDO are included since they are considered to be possible climate drivers for rainfall – as we now explain.

The world's oceans have the largest heat capacity of any single component of the climate system, and over the past 40 years they have been the dominant source of changes in global heat content (Levitus *et al.* 2001; Willis, Roemmich & Cornuelle 2004). We used the time series of three-monthly OHA, measured in 10^{22} joules, for the 0–700 m layer in the Southern Pacific Ocean Basin. At the time of writing, these data are available from the US National Oceanographic and Data Center web-site www.nodc.noaa.gov/OC5/3M_HEAT_CONTENT. In September 2010, the OHA data were updated, and we used this new version in the analysis.

The Southern Oscillation Index (SOI) is calculated from the monthly or seasonal fluctuations in the air pressure difference between Tahiti and Darwin, Australia, and is a unitless quantity. Monthly values are available from the Australian Bureau of Meteorology (current web-site: www.bom.gov.au/climate/current/soihtm1.shtml). Sustained negative values of the SOI often indicate El Niño episodes, resulting in a reduction in rainfall over eastern and northern Australia. Positive values of the SOI are associated with an increased probability that eastern and northern Australia will be wetter than normal.

Monthly PDO data are currently available from the web-site jisao.washington.edu/pdo/PDO.latest. There the data are described as being 'the leading principal component

of monthly sea surface temperature anomalies in the North Pacific Ocean, poleward of 20 degrees North. The monthly mean global average sea surface temperature anomalies are removed to separate this pattern of variability from any "global warming" signal that may be present in the data.' Note that PDO is also a unitless quantity and is a long-lived El Niño-like measure of Pacific climate variability. While SOI and PDO have similar spatial climatic fingerprints, they have very different behaviours in time. Hence, PDO is included to examine whether it can explain further variability in extreme rainfall.

The following GEV geoadditive model was fitted via Algorithm 3.3:

$$\begin{aligned}\texttt{winter max. rainfall}_i \overset{\text{ind.}}{\sim} \text{GEV}(&f_1(\texttt{year}_i) + f_2(\texttt{day in season}_i) + f_3(\texttt{OHA}_i) \\ &+ f_4(\texttt{SOI}_i) + f_5(\texttt{PDO}_i) + g(\texttt{longitude}_i, \texttt{latitude}_i), \sigma_\varepsilon, \xi) \quad (15)\end{aligned}$$

for $1 \leq i \leq n$, where $n = 1874$ is the total number of winter maximum rainfall measurements from the 50 weather stations from Figure 1 between the years 1955 and 2003 (not all stations had this full set of years). Each univariate function estimate used 37 spline basis functions of the type described in Wand & Ormerod (2008). The bivariate function estimate used 50 bivariate thin-plate spline basis functions, as described in section 13.5 of Ruppert, Wand & Carroll (2003), with knots at the weather stations. The hyperparameters were set at:

$$\begin{aligned}&\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = 10^8\, \boldsymbol{I};\ A_\varepsilon = B_\varepsilon = A_u = B_u = 0.01; \\ &p(\xi) \text{ uniform on } \ \Xi = \{0.00, 0.01, \ldots, 0.50\}.\end{aligned} \quad (16)$$

To ensure scale invariance, all variables were standardized prior to input into Algorithm 3.3. The results were then back-transformed to the original units. These hyperparameter settings correspond to a low amount of prior information.

The smooth functions of year ($f_1(\texttt{year}_i)$) and geographical location ($g(\texttt{longitude}_i, \texttt{latitude}_i)$) account for anticipated temporal and spatial correlation in the winter maximum rainfalls. The former of these de-correlation devices is commonly used in generalized additive model analysis of environmental epidemiologic time series data (e.g. Wand & Schwartz 2002; Dominici, McDermott & Hastie 2004).

Figures 3 and 4 show, respectively, the estimated univariate functions and bivariate function from MFVB fitting of (15) with hyperparameters (16). The plotting scheme described in Section 4 is used in Figure 3.

The smooth function of year shows pronounced oscillation– corresponding to drought and wet periods in the Sydney hinterland. The fitted surface for geographical location reflects well-known geographical patterns such as higher rainfall along the New South Wales coastal plain and orographic effects due to the Great Dividing Range. The effect of OHA is weakly nonlinear up to OHA $= 1.7$. However, for OHA ≥ 1.7 there is a sharp upward ramp. SOI has an interesting approximate piecewise linear effect. For SOI increasing up to about SOI $= 10$ there is a positive effect on maximum rainfall. This is consistent with the aforementioned association between high SOI and wet weather in eastern Australia. However, higher SOI values tend to have a negative effect. The estimate of $f_5(\texttt{PDO})$ shows an interesting oscillatory relationship.

The posterior probability mass function for the GEV shape parameter, ξ, is shown in Figure 5. Most of the probability mass is between 0.15 and 0.27, with the posterior mode at

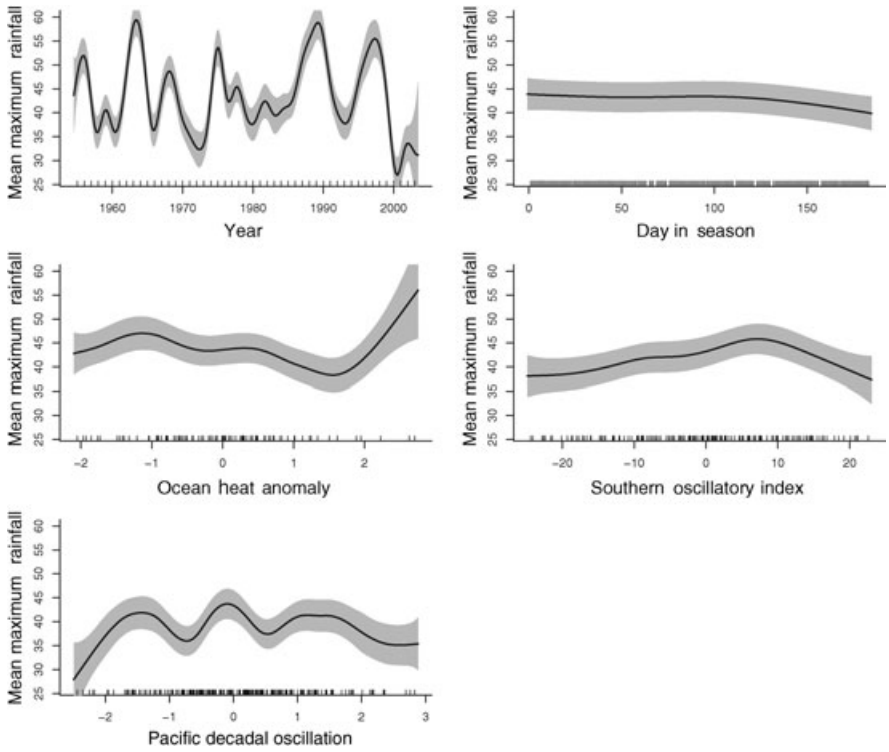

Figure 3. MFVB univariate functional fits in the GEV additive model (15) for the Sydney hinterland rainfall data. The vertical axis in each panel is such that all additive functions are included, with the horizontal axis predictor varying and the other predictors set at their average values, as decribed in Section 4. The grey region corresponds to approximate pointwise 95% credible sets.

$\xi = 0.21$. This is consistent with table 4 of Koutsoyiannis (2004), from which an approximate 95% confidence interval of $(-0.054, 0.26)$ for ξ can be derived after averaging maximum rainfall estimates of individual GEV fits to the annual maximum daily rainfall series for 169 data sets from Europe and USA.

7. Comparisons with Markov Chain Monte Carlo

As mentioned in Section 1, speed is the main attraction of MFVB when compared with MCMC. Our R program for performing the analysis of the Sydney hinterland rainfall data, as described in Section 6, takes about 4.5 minutes to run on the third author's laptop (Mac OS X; 2.66-GHz processor, 43 GBytes of random access memory). Such speed is quite important in the model development phase. On the other hand, MCMC implementation of the same model via the R package BRUGS (Ligges *et al.* 2009), and with 10,000 MCMC iterations, took just over 21 hours to run on the same computer. Hence, 'off-the-shelf' MCMC-based GEV additive model analyses can be quite difficult owing to the long waiting period between model fits. Faster MCMC implementations are possible, but would require a great deal of additional programming.

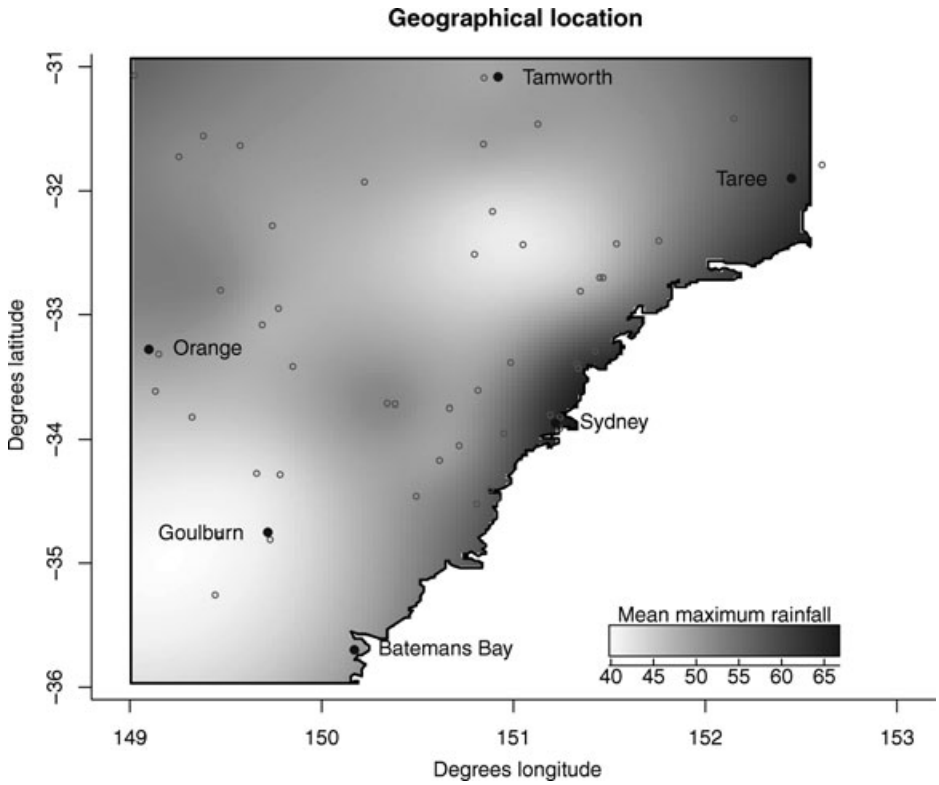

Figure 4. MFVB bivariate functional fit for geographical location in the GEV additive model (15) for the Sydney hinterland rainfall data. The weather station locations are shown as grey dots. The black dots show the locations of six cities and towns with names as labelled.

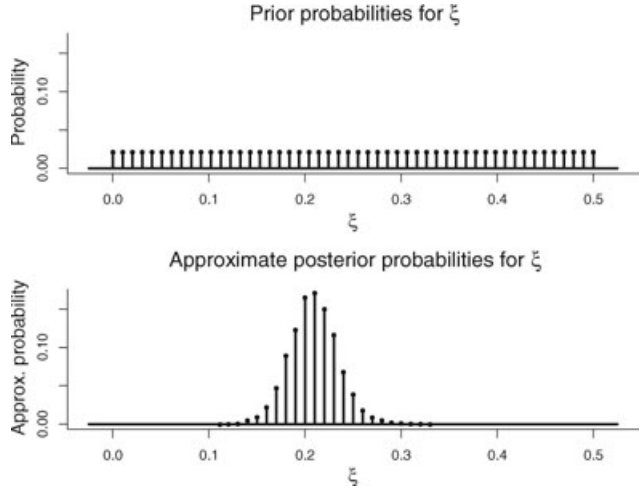

Figure 5. The prior and MFVB approximate posterior probability mass functions for the GEV shape parameter ξ in the GEV additive model (15) for the Sydney hinterland maximum rainfall data.

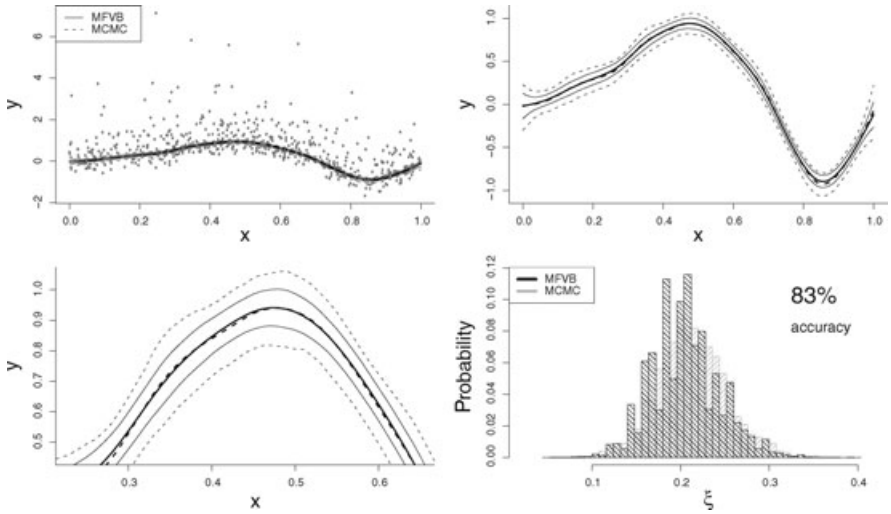

Figure 6. Accuracy comparison between MFVB and MCMC for a single predictor model ($d = 1$). Top left panel: the simulated data together with the MFVB-based and MCMC-based estimates of f and pointwise 95% credible sets. Top right panel: the same as the top left panel but without the data and with the frame modified to zoom in on the function estimates. Bottom left panel: the same as the top right panel, but with the frame modified to zoom in on the region surrounding the peaks of the two function estimates. Bottom right panel: posterior probability mass functions for ξ based on MFVB and MCMC fitting. The accuracy measure is based on the summed absolute difference between the probabilities such that 100% accuracy corresponds to equality.

We have done some cursory accuracy comparisons between MFVB and MCMC. Figure 6 provides an illustrative summary for the single predictor ($d = 1$) case. The sample size is $n = 500$, and the data were simulated according to $y_i \overset{\text{ind.}}{\sim} \text{GEV}(f(x_i), 0.5, 0.3)$ with $f(x) = \sin(2\pi x^2)$. The x_i values were generated from the uniform distribution on $(0, 1)$. The hyperparameters were taken to be the same as those give in (16) but with $\Xi = \{0.00, 0.08, \ldots, 0.40\}$. The MCMC fit was performed using BRUGS with a burn-in of size 5000, and a thinning factor of 5 applied to the subsequent 5000 MCMC samples.

The first three panels show MFVB-based and MCMC-based estimates of f, together with pointwise 95% credible sets. The top right and bottom left panels omit the data and show zoomed versions of the estimates to allow better comparison. The function estimates are seen to be very close. However, the pointwise credible intervals differ substantially. In particular, those based on MFVB are overly narrow. This observation is in keeping with those made by Wand *et al.* (2011) for simpler GEV models. This behaviour is typical of all the MFVB versus MCMC comparisons we have performed for GEV additive models. It suggests that MFVB leads to accurate recovery of the mean structure, but that the credible interval bands are not suitable for valid pointwise inference for the mean function. Nevertheless, regardless of inferential benefits, variability bands are a useful complement to curve estimates since they convey other aspects such as influence.

As shown in the bottom right panel of Figure 6, the accuracy of MFVB for inference concerning ξ is quite good. In this example, MFVB achieves an accuracy of 83% for $p(\xi \mid \boldsymbol{y})$,

where the accuracy is defined as a percentage given by

$$100\left\{1-\frac{1}{2}\sum_{\xi\in\Xi}|\,q^*(\xi)-\widehat{p}(\xi\,|\,\boldsymbol{y})|\right\},$$

where $\widehat{p}(\xi\,|\,\boldsymbol{y})$ is the approximation to $p(\xi\,|\,\boldsymbol{y})$ based on the MCMC sample. It is easy to verify that this accuracy measure is non-negative and equals 100% if and only if the two probability mass functions coincide exactly. A larger simulation study in Wand *et al.* (2011) corroborates the finding that MFVB is accurate for the GEV shape parameter.

8. Discussion

We have developed a new method for GEV additive model analysis. Comparison studies suggest that the MFVB estimation of the additive model components and shape parameter is highly accurate, but that the credible sets are overly narrow. Nevertheless, it facilitates approximate Bayesian inference for such analysis in a fraction of the time taken by MCMC, and therefore has viability advantages for larger models and sample sizes.

Appendix: Derivations

This appendix contains derivations of the optimal q densities. The parameters of these form the basis of Algorithm 3.2. We also give the derivation of $\log\,\underline{p}(\boldsymbol{y};q)$ given in (11).

A.1. Full conditionals

We use 'rest' to denote the random components of the model not including the current argument. Additive constants with respect to the function argument are denoted by 'const.'.

Define

$$\boldsymbol{\Omega}=\boldsymbol{C}^\top\left(\frac{1}{\sigma_\varepsilon^2}\sum_{k=1}^{\mathcal{K}}\frac{1}{s_k^2}\boldsymbol{D}_{\boldsymbol{a}_k}\right)\boldsymbol{C}+\text{blockdiag}\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1},\frac{1}{\sigma_{u1}^2}\boldsymbol{I}_{K_1},\ldots,\frac{1}{\sigma_{ud}^2}\boldsymbol{I}_{K_d}\right)$$

$$\text{and}\quad\boldsymbol{\omega}=\boldsymbol{C}^\top\left\{\frac{1}{\sigma_\varepsilon^2}\sum_{k=1}^{\mathcal{K}}\frac{1}{s_k^2}\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y}-\sigma_\varepsilon m_k\mathbf{1})\right\}.$$

Then the full conditionals satisfy

$$\log\,p(\boldsymbol{\beta},\boldsymbol{u}|\text{rest})=-\frac{1}{2}\left[\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^\top\boldsymbol{\Omega}\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}-2\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}^\top\boldsymbol{\omega}\right]+\text{const.},$$

$$\begin{aligned}\log p\big(\sigma_\varepsilon^2\big|\text{rest}\big)=&-\left(A_\varepsilon+\frac{1}{2}n+1\right)\log\big(\sigma_\varepsilon^2\big)\\&-\frac{1}{2\sigma_\varepsilon^2}\sum_{k=1}^{\mathcal{K}}\frac{1}{s_k^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})^\top\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})\\&-\frac{B_\varepsilon}{\sigma_\varepsilon^2}+\frac{1}{\sigma_\varepsilon}\sum_{k=1}^{\mathcal{K}}\frac{m_k}{s_k^2}\mathbf{1}^\top\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})+\text{const.},\end{aligned}$$

$$\log\ p\big(\sigma_{u1}^2,\ldots,\sigma_{ud}^2|\text{rest}\big)=\sum_{\ell=1}^{d}\Bigg[-\left(A_{u\ell}+\frac{1}{2}K_\ell+1\right)\log\big(\sigma_{u\ell}^2\big)$$
$$-\left(B_{u\ell}+\frac{1}{2}\|\boldsymbol{u}_\ell\|^2\right)\Big/\sigma_{u\ell}^2\Bigg]+\text{const.},$$
$$\log\ p(\boldsymbol{a}|\text{rest})=\sum_{i=1}^{n}\sum_{k=1}^{\mathcal{K}}a_{ik}\left\{\log(w_k/s_k)-\frac{\{(\boldsymbol{y}-\boldsymbol{X\beta}-\boldsymbol{Zu})_i-\sigma_\varepsilon m_k\}^2}{2\sigma_\varepsilon^2 s_k^2}\right\}+\text{ const.}$$

These expressions follow from standard manipulations, reminiscent of those used in the MCMC literature. Note that the graphical representation of the model given in Figure 2 and the theory of Markov blankets (Pearl 1988) is helpful in the derivations. For example,

$$p\big(\sigma_{u\ell}^2|\text{rest}\big)=p\big(\sigma_{u\ell}^2|\text{Markov blanket of }\sigma_{u\ell}^2\big)=p\big(\sigma_{u\ell}^2|\beta_\ell,\boldsymbol{u}_\ell\big).\tag{17}$$

The Markov blanket of a node in a directed acyclic graph consists of its parents, children and co-parents.

A.2. Optimal q densities

A.2.1.

Expression for $q^*(\boldsymbol{\beta},\boldsymbol{u})$

$$q^*(\boldsymbol{\beta},\boldsymbol{u})\sim\text{N}(\mu_{q(\boldsymbol{\beta},\boldsymbol{u})},\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{\mu})})$$

where

$$\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}=\boldsymbol{C}^\top\left(\mu_{q(1/\sigma_\varepsilon^2)}\sum_{k=1}^{\mathcal{K}}\frac{1}{s_k^2}\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\right)\boldsymbol{C}+\text{blockdiag}\big(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1},\mu_{q(1/\sigma_{u1}^2)}\boldsymbol{I}_{K_1},\ldots,\mu_{q(1/\sigma_{ud}^2)}\boldsymbol{I}_{K_d}\big)$$

and

$$\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{\mu})}=\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\left(\mu_{q(1/\sigma_\varepsilon^2)}\sum_{k=1}^{\mathcal{K}}\frac{1}{s_k^2}\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{y}-\mu_{q(1/\sigma_\varepsilon)}\sum_{k=1}^{\mathcal{K}}\frac{m_k}{s_k^2}\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{1}\right).$$

Derivation:

$$\log\ q^*(\boldsymbol{\beta},\boldsymbol{u})=E_q\{\log\ p(\boldsymbol{\beta},\boldsymbol{u}|\text{rest})\}+\text{const.}$$
$$=-\frac{1}{2}\left\{\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{u}\end{bmatrix}^\top E_q(\boldsymbol{\Omega})\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{u}\end{bmatrix}-2\begin{bmatrix}\boldsymbol{\beta}\\\boldsymbol{u}\end{bmatrix}^\top E_q(\boldsymbol{\omega})\right\}+\text{const.}$$

The result then follows after taking expectations of $\boldsymbol{\Omega}$ and $\boldsymbol{\omega}$ with respect to the q densities, and identification of the multivariate normal mean and covariance matrix via 'completion of the square' manipulations.

A.2.2.

Expressions for $q^*(\sigma_\varepsilon^2)$ and $\mu_{q(1/\sigma_\varepsilon^2)}$, $\mu_{q(1/\sigma_\varepsilon)}$

$$q^*\big(\sigma_\varepsilon^2\big) = \frac{\big(\sigma_\varepsilon^2\big)^{-A-\frac{1}{2}n-1}\exp\big\{C_9\big/\big(\sigma_\varepsilon^2\big)^{1/2} - C_{10}/\sigma_\varepsilon^2\big\}}{2\mathcal{J}^+(2A+n-1, C_9, C_{10})}, \quad \sigma_\varepsilon^2 > 0,$$

$$\mu_{q(1/\sigma_\varepsilon^2)} = \frac{\mathcal{J}^+(2A+n+1, C_9, C_{10})}{\mathcal{J}^+(2A+n-1, C_9, C_{10})} \quad \text{and} \quad \mu_{q(1/\sigma_\varepsilon)} = \frac{\mathcal{J}^+(2A+n, C_9, C_{10})}{\mathcal{J}^+(2A+n-1, C_9, C_{10})}$$

where

$$C_9 = \sum_{k=1}^{K} \frac{m_k}{s_k^2}\mathbf{1}^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})$$

$$\text{and} \quad C_{10} = B_\varepsilon + \frac{1}{2}\sum_{k=1}^{K}\frac{1}{s_k^2}\big\{\text{tr}\,\big(\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{C}\big) + (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})\big\}.$$

Derivations:

First,

$$\begin{aligned}\log\, q^*\big(\sigma_\varepsilon^2\big) = &-\left(A_\varepsilon + \frac{1}{2}n + 1\right)\log\big(\sigma_\varepsilon^2\big)\\ &-\frac{1}{2\sigma_\varepsilon^2}\sum_{k=1}^{K}\frac{1}{s_k^2}E_q\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^\top \boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\}\\ &-\frac{B_\varepsilon}{\sigma_\varepsilon^2} + (1/\sigma_\varepsilon)\sum_{k=1}^{K}\frac{m_k}{s_k^2}\mathbf{1}^\top E_q\{\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\} + \text{const.}\end{aligned}$$

Also,

$$\begin{aligned}&E_q\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})^\top \boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\}\\ &\quad = E_q\big[\text{tr}\big\{\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y} - \boldsymbol{C}[\boldsymbol{\beta}^\top\ \boldsymbol{u}^\top]^\top)(\boldsymbol{y} - \boldsymbol{C}[\boldsymbol{\beta}^\top\ \boldsymbol{u}^\top]^\top)^\top\big\}\big]\\ &\quad = \text{tr}\big[\boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\big\{\big(\boldsymbol{y} - \boldsymbol{C}\big[\boldsymbol{\mu}_{q(\boldsymbol{\beta})}{}^\top\ \boldsymbol{\mu}_{q(\boldsymbol{u})}{}^\top\big]^\top\big)\big(\boldsymbol{y} - \boldsymbol{C}\big[\boldsymbol{\mu}_{q(\boldsymbol{\beta})}{}^\top\ \boldsymbol{\mu}_{q(\boldsymbol{u})}{}^\top\big]^\top\big)^\top + \text{cov}_q(\boldsymbol{C}[\boldsymbol{\beta}^\top\ \boldsymbol{u}^\top])\big\}\big]\\ &\quad = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})}) + \text{tr}\big\{\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}\boldsymbol{C}\big\}.\end{aligned}$$

Similarly, $E_q\{\boldsymbol{D}_{\boldsymbol{a}_k}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})\} = \boldsymbol{D}_{\boldsymbol{\mu}_{q(\boldsymbol{a}_k)}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\,\boldsymbol{\mu}_{q(\boldsymbol{u})})$. It follows that

$$q^*\big(\sigma_\varepsilon^2\big) \propto \big(\sigma_\varepsilon^2\big)^{-A-\frac{1}{2}n-1}\exp\big\{C_9/\big(\sigma_\varepsilon^2\big)^{1/2} - C_{10}\big/\sigma_\varepsilon^2\big\}, \quad \sigma_\varepsilon^2 > 0.$$

Standard manipulations lead to $2\mathcal{J}^+(2A+n-1, C_9, C_{10})$ being the normalizing factor. The expressions for $\mu_{q(1/\sigma_\varepsilon^2)}$ and $\mu_{q(1/\sigma_\varepsilon)}$ involve similar manipulations.

A.2.3.

Expressions for $q^*(\sigma_{u1}^2, \ldots, \sigma_{ud}^2)$ and $\mu_{q^*(1/\sigma_{u\ell}^2)}$, $1 \le \ell \le d$

$q^*(\sigma_{u1}^2, \ldots, \sigma_{ud}^2)$ is a product of dIG$(A_{u\ell} + \frac{1}{2}K_\ell, B_{q(\sigma_{u\ell}^2)})$ density functions, where

$$B_{q(\sigma_{u\ell}^2)} = B_{u\ell} + \frac{1}{2}\big\{\|\boldsymbol{\mu}_{q(\mu_\ell)}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\big\}.$$

Also,

$$\mu_{q^*(1/\sigma_{u\ell}^2)} = \left(A_{u\ell} + \frac{1}{2}K_\ell\right) \Big/ B_{q(\sigma_{u\ell}^2)},\ 1 \le \ell \le d.$$

Derivations:

$$\begin{aligned}\log q^*\big(\sigma_{u1}^2, \ldots, \sigma_{ud}^2\big) &= E_q\{\log\ p(\sigma_{u1}^2, \ldots, \sigma_{ud}^2|\text{rest})\} + \text{const.}\\ &= \sum_{\ell=1}^{d}\left[-\left(A_{u\ell} + \frac{1}{2}K_\ell + 1\right)\log\big(\sigma_{u\ell}^2\big) - \left\{B_{u\ell} + \frac{1}{2}E_q\big(\|\boldsymbol{u}_\ell\|^2\big)\right\}\Big/\sigma_{u\ell}^2\right]\\ &\quad + \text{const.}\end{aligned}$$

Noting that $E_q(\|\boldsymbol{u}_\ell\|^2) = \|\boldsymbol{\mu}_{q(\mu_\ell)}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})$ we have

$$q^*\big(\sigma_{u1}^2, \ldots, \sigma_{ud}^2\big) \propto \prod_{\ell=1}^{d}\big(\sigma_{u\ell}^2\big)^{-A_{u\ell}-\frac{1}{2}K_\ell-1}\exp\big\{-B_{q(\sigma_{u\ell}^2)}\big/\sigma_{u\ell}^2\big\},$$

and the stated result for $q^*(\sigma_{u1}^2, \ldots, \sigma_{ud}^2)$ follows after normalization. The expression for the $\mu_{q^*(1/\sigma_{u\ell}^2)}$ follows from standard manipulations involving inverse gamma density functions.

A.2.4.

Expressions for $q^*(\boldsymbol{a})$ and $\mu_{q(a_{ik})}$

$$q^*(\boldsymbol{a}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\{\mu_{q(a_{ik})}\}^{a_{ik}} \quad \text{where} \quad \mu_{q(a_{ik})} = \exp(\nu_{ik})\Big/\sum_{k=1}^{\mathcal{K}}\exp(\nu_{ik})$$

and

$$\begin{aligned}\nu_{ik} = \log(w_k/s_k) - \frac{1}{2s_k^2}\big[&\mu_{q(1/\sigma_\varepsilon^2)}\big\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})_i^2 + \big(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top\big)_{ii}\big\}\\ &- 2\mu_{q(1/\sigma_\varepsilon)}m_k(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \boldsymbol{Z}\boldsymbol{\mu}_{q(\boldsymbol{u})})_i + m_k^2\big].\end{aligned}$$

Derivation:

$$\log\ q^*(\boldsymbol{a}) = \sum_{i=1}^{n}\sum_{k=1}^{\mathcal{K}}a_{ik}\left(\log(w_k/s_k) - E_q\left[\frac{\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i - \sigma\, m_k\}^2}{2\sigma^2 s_k^2}\right]\right) + \text{const.}$$

Note that

$$\frac{\{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i - \sigma m_k\}^2}{2\sigma^2 s_k^2} = \frac{1}{2s_k^2}\left\{\sigma_\varepsilon^{-2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i^2 - 2\sigma_\varepsilon^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{Z}\boldsymbol{u})_i\, m_k + m_k^2\right\}$$

and also that

$$\begin{aligned} E_q\{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}-\boldsymbol{Z}\boldsymbol{u})_i^2\} &= \{\boldsymbol{y}-\boldsymbol{X}\,E_q(\boldsymbol{\beta})-\boldsymbol{Z}\,E_q(\boldsymbol{u})\}_i^2+\mathrm{var}_q\{(\boldsymbol{C}[\boldsymbol{\beta}^\top\ \boldsymbol{u}^\top]^\top)_i\} \\ &= \{\boldsymbol{y}-\boldsymbol{X}\,\mu_{q(\boldsymbol{\beta})}-\boldsymbol{Z}\,\mu_{q(\boldsymbol{u})}\}_i^2+\{\mathrm{cov}_q(\boldsymbol{C}[\boldsymbol{\beta}^\top\ \boldsymbol{u}^\top]^\top)\}_{ii} \\ &= \{\boldsymbol{y}-\boldsymbol{X}\,\mu_{q(\boldsymbol{\beta})}-\boldsymbol{Z}\,\mu_{q(\boldsymbol{u})}\}_i^2+(\boldsymbol{C}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta},\boldsymbol{u})}\boldsymbol{C}^\top)_{ii}. \end{aligned}$$

The expression for v_{ik} follows immediately. The expression for $\mu_{q(a_{ik})}$ then follows from the requirement that $\sum_{k=1}^{K}\mu_{q(a_{ik})}=1$.

A.2.5.

Expression for $\log\{\underline{p}(\boldsymbol{y};q)\}$

Derivation:

$$\begin{aligned} \log\,\underline{p}(\boldsymbol{y};q) &= E_q\{\log\,p(\boldsymbol{y},\boldsymbol{a},\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\sigma_{u1}^2,\ldots,\sigma_{ud}^2)-\log\,q^*(\boldsymbol{a},\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\sigma_{u1}^2,\ldots,\sigma_{ud}^2)\} \\ &= E_q\{\log\,p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\boldsymbol{a})\}+E_q\{\log\,p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_{u1}^2,\ldots,\sigma_{ud}^2)-\log\,q^*(\boldsymbol{\beta},\boldsymbol{u})\} \\ &\quad +E_q\{\log p(\boldsymbol{a})-\log q^*(\boldsymbol{a})\}+E_q\{\log\,p(\sigma_\varepsilon^2)-\log\,q^*(\sigma_\varepsilon^2)\} \\ &\quad +E_q\{\log\,p(\sigma_{u1}^2,\ldots,\sigma_{ud}^2)-\log\,q^*(\sigma_{u1}^2,\ldots,\sigma_{ud}^2)\}. \end{aligned}$$

First,

$$\begin{aligned} \log\,\{p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\boldsymbol{a})\} &= -\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma_\varepsilon^2)-\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}\log(s_k^2) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{a_{ik}\{y_i-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i-\sigma_\varepsilon m_k\}^2}{\sigma_\varepsilon^2}. \end{aligned}$$

Then, noting that

$$\frac{\{y_i-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i-\sigma_\varepsilon m_k\}^2}{\sigma_\varepsilon^2}=\frac{\{y_i-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i\}^2}{\sigma_\varepsilon^2}-\frac{2m_k\{y_i-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u})_i\}}{\sigma_\varepsilon}+m_k^2,$$

we have

$$\begin{aligned} E_q\{\log\,p(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma_\varepsilon^2,\boldsymbol{a})\} &= -\frac{n}{2}\log(2\pi)-\frac{n}{2}E_q\{\log(\sigma_\varepsilon^2)\}+C_9\mu_{q(1/\sigma_\varepsilon)} \\ &\quad -(C_{10}-B_\varepsilon)\mu_{q(1/\sigma_\varepsilon^2)}-\frac{1}{2}\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}\{\log(s_k^2)+(m_k^2/s_k^2)\}. \end{aligned}$$

Second,

$$\log p(\boldsymbol{a})-\log q^*(\boldsymbol{a})=\sum_{i=1}^{n}\sum_{k=1}^{K}a_{ik}(\log w_k-\log\mu_{q(a_{ik})}),$$

and so

$$E_q\{\log p(\boldsymbol{a})-\log q^*(\boldsymbol{a})\}=\sum_{k=1}^{K}\mu_{q(a_{\bullet k})}\log(w_k)-\sum_{i=1}^{n}\sum_{k=1}^{K}\mu_{q(a_{ik})}\log(\mu_{q(a_{ik})}).$$

Third,

$$\log\ p\big(\sigma_\varepsilon^2\big) - \log\ q^*\big(\sigma_\varepsilon^2\big) = A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) + \frac{n}{2}\log\big(\sigma_\varepsilon^2\big) + (C_9 - B_\varepsilon)/\sigma_\varepsilon^2 - C_{10}\big/\big(\sigma_\varepsilon^2\big)^{1/2} + \log\mathcal{J}^+(2A_\varepsilon + n - 1, C_9, C_{10}, 0) + \log(2),$$

which leads to

$$E_q\big\{\log\ p\big(\sigma_\varepsilon^2\big) - \log\ q^*\big(\sigma_\varepsilon^2\big)\big\} = A_\varepsilon \log(B_\varepsilon) - \log\Gamma(A_\varepsilon) + \frac{n}{2}E_q\big\{\log\big(\sigma_\varepsilon^2\big)\big\} + (C_{10} - B_\varepsilon)\mu_{q(1/\sigma_\varepsilon^2)} - C_9\mu_{q(1/\sigma_\varepsilon)} + \log\mathcal{J}^+(2A_\varepsilon + n - 1, C_9, C_{10}) + \log(2).$$

Next,

$$\begin{aligned} E_q\big\{\log\ p\big(\boldsymbol{\beta}, \boldsymbol{u}|\sigma_{u1}^2, \ldots, \sigma_{ud}^2\big)\big\} &= -\frac{1}{2}\left(1 + d + \sum_{\ell=1}^{d} K_\ell\right)\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \\ &\quad - \sum_{\ell=1}^{d}\frac{1}{2}K_\ell E_q\big\{\log\big(\sigma_{u\ell}^2\big)\big\} - \frac{1}{2}E_q\big(\boldsymbol{\beta}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\beta}\big) \\ &\quad - \frac{1}{2}\sum_{\ell=1}^{d} E_q\big(1\big/\sigma_{u\ell}^2\big)\, E_q\big(\|\boldsymbol{u}_\ell\|^2\big) \\ &= \frac{1}{2}\left(p + \sum_{\ell=1}^{d} K_\ell\right)\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_{\boldsymbol{\beta}}| \\ &\quad - \sum_{\ell=1}^{d}\frac{1}{2}K_\ell E_q\big\{\log\big(\sigma_{u\ell}^2\big)\big\} \\ &\quad - \frac{1}{2}\boldsymbol{\mu}_{q(\boldsymbol{\beta})}{}^\top\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{q(\boldsymbol{\beta})} - \frac{1}{2}\mathrm{tr}\big(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\Sigma}_{q(\boldsymbol{\beta})}\big) \\ &\quad - \frac{1}{2}\sum_{\ell=1}^{d}\mu_{q(1/\sigma_{u\ell}^2)}\big\{\|\mu_{q(\boldsymbol{u}_\ell)}\|^2 + \mathrm{tr}\,(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\big\} \end{aligned}$$

and

$$-E_q\{\log\ q(\boldsymbol{\beta}, \boldsymbol{u})\} = \frac{1}{2}\left(1 + d + \sum_{\ell=1}^{d} K_\ell\right)\{1 + \log(2\pi)\} + \frac{1}{2}\log|\boldsymbol{\Sigma}_{q(\beta,\boldsymbol{u})}|.$$

Finally,

$$\begin{aligned} &E_q\big\{\log\ p\big(\sigma_{u1}^2, \ldots, \sigma_{ud}^2\big) - \log\ q^*\big(\sigma_{u1}^2, \ldots, \sigma_{ud}^2\big)\big\} \\ &\quad = \sum_{\ell=1}^{d}\big[A_{u\ell}\log(B_{u\ell}) - \log\Gamma(A_{u\ell}) - (A_{u\ell} + 1)E_q\big\{\log\big(\sigma_{u\ell}^2\big)\big\} - B_{u\ell}E_q\big(1\big/\sigma_{u\ell}^2\big) \\ &\qquad - A_{q(\sigma_{u\ell}^2)}\log\big(B_{q(\sigma_{u\ell}^2)}\big) + \log\Gamma\big(A_{q(\sigma_{u\ell}^2)}\big) + \big(A_{q(\sigma_{u\ell}^2)} + 1\big)E_q\big\{\log\big(\sigma_{u\ell}^2\big)\big\} + B_{q(\sigma_{u\ell}^2)}E_q\big(1/\sigma_{u\ell}^2\big)\big] \end{aligned}$$

$$= \sum_{\ell=1}^{d} \left[A_{u\ell} \log(B_{u\ell}) - \log \Gamma(A_{u\ell}) - \left(A_{u\ell} + \frac{1}{2} K_\ell \right) \log \left(B_{q(\sigma_{u\ell}^2)} \right) + \log \Gamma \left(A_{u\ell} + \frac{1}{2} K_\ell \right) \right.$$
$$\left. + \frac{1}{2} K_\ell E_q \left\{ \log \left(\sigma_{u\ell}^2 \right) \right\} + \mu_{q(1/\sigma_{u\ell}^2)} \left(B_{q(\sigma_{u\ell}^2)} - B_{u\ell} \right) \right].$$

Combining each of the terms, and noting the relationship $B_{q(\sigma_{u\ell}^2)} = B_{u\ell} + \frac{1}{2}\{\|\boldsymbol{\mu}_{q(\mu_\ell)}\|^2 + \operatorname{tr}(\boldsymbol{\Sigma}_{q(\boldsymbol{u}_\ell)})\}$, $1 \le \ell \le d$, from the Algorithm 3.2 updates, we obtain (11).

References

ATTIAS, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence*, 21–30. San Francisco: Morgan Kaufmann.

BISHOP, C.M. (2006). *Pattern Recognition and Machine Learning*. New York: Springer.

BRAUN, M. & MCAULIFFE, J. (2010). Variational inference for large-scale models of discrete choice. *J. Amer. Statist. Assoc.* **106**, 324–335.

CHAVEZ-DEMOULIN, V. & DAVISON, A.C. (2005). Generalized additive modelling of sample extremes. *J. R. Statist. Soc. Ser. C* **54**, 207–222.

CRAINICEANU, C., RUPPERT, D. & WAND, M.P. (2005). Bayesian analysis for penalized spline regression using WinBUGS. *J. Statist. Softw.* Vol. 14, Article 14, 1–24.

DAVISON, A.C. & RAMESH, N.I. (2000). Local likelihood smoothing of sample extremes. *J. R. Statist. Soc. Ser. B* **62**, 191–208.

DOMINICI, F., MCDERMOTT, A. & HASTIE, T. (2004). Improved semiparametric time series models of air pollution and mortality. *J. Amer. Statist. Assoc.* **99**, 938–948.

FAES, C., ORMEROD, J.T. & WAND, M.P. (2011). Variational Bayesian inference for parametric and nonparametric regression with missing data. *J. Amer. Statist. Assoc.* **106**, 959–971.

GRADSHTEYN, I.S. & RYZHIK, I.M. (1994). *Tables of Integrals, Series, and Products*, 5th edn. San Diego, CA: Academic Press.

GURRIN, L.C., SCURRAH, K.J. & HAZELTON, M.L. (2005). Tutorial in biostatistics: spline smoothing with linear mixed models. *Statist. Med.* **24**, 3361–3381.

HENNERFEIND, A., BREZGER, A. & FAHRMEIR, L. (2006). Geoadditive survival models. *J. Amer. Statist. Assoc.*, **101**, 1065–1075.

JEFFREY, S.J., CARTER, J.O., MOODIE, K.M. & BESWICK, A.R. (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environ. Modell. Softw.* **16**, 309–330.

KAMMANN, E.E. & WAND, M.P. (2003). Geoadditive models. *J. R. Statist. Soc. Ser. C* **52**, 1–18.

KOUTSOYIANNIS, D. (2004). Exploration of long records of extreme rainfall and design rainfall inferences. In *Hydrology: Science & Practice for the 21st Century*, eds. B. Webb, N. Arnell, C. Onof, N. MacIntire, R. Gurney and C. Kirby, vol. 1, pp. 148–157. London: British Hydrological Society.

LAURINI, F. & PAULI, F. (2009). Smoothing sample extremes: the mixed model approach. *Comput. Statist. Data Analysis*, **53**, 3842–3854.

LEVITUS, S., ANTONOV, J.I., WANG, J., DELWORTH, T.L., DIXON, K.W. & BROCCOLI, A.J. (2001). Anthropogenic warming of Earth's climate system. *Science* **292**, 267–270.

LIGGES, U., THOMAS, A., SPIEGELHALTER, D., BEST, N. LUNN, D., RICE, K. & STURTZ, S. (2009). BRugs 0.5: OpenBUGS and its R/S-PLUS interface BRugs. Available from URL: http://www.stats.ox.ac.uk/pub/RWin/src/contrib/

LUENBERGER, D.G. & YE, Y. (2008). *Linear and Nonlinear Programming*, 3rd edn. New York: Springer.

MARLEY, J.K. & WAND, M.P. (2010). Non-standard semiparametric regression via BRugs. *J. Statist. Softw.* **37**(5), 1–30.

ORMEROD, J.T. & WAND, M.P. (2010). Explaining variational approximations. *Amer. Statist.*, **64**, 140–153.

PADOAN, S.A. & WAND, M.P. (2008). Mixed-model based additive models for sample extremes. *Statist. Probab. Lett.* **78**, 2850–2858.

PEARL, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufmann.

R Development Core Team (2010). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available from URL: http://www.R-project.org

REN, Q., BANERJEE, S., FINLEY, A.O. & HODGES, J.S. (2011). Variational Bayesian methods for spatial data analysis. *Comput. Statist. Data Analysis* **55**, 3197–3217.

RUPPERT, D., WAND, M.P. & CARROLL, R.J. (2003). *Semiparametric Regression*. New York: Cambridge University Press.

SAUL, L.K. & JORDAN, M.I. (1996). Exploiting tractable substructures in intractable networks. In *Advances in Neural Information Processing Systems*, pp. 435–442, Cambridge, MA: MIT Press.

WAHBA, G., WANG, Y., GU, C., KLEIN, R. & KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsiin epidemiological study of diabetic retinopathy. *Annals Statist.* **23**, 1865–1895.

WAND, M.P. & ORMEROD, J.T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Aust. N. Z. J. Stat.* **50**, 179–198.

WAND, M.P. & SCHWARTZ, J. (2002). Smoothing in environmental epidemiology. *Encyclopedia of Environmetrics* **4**, 2020–2023.

WAND, M.P., ORMEROD, J.T., PADOAN, S.A. & FRÜHWIRTH, R. (2011). Mean field variational Bayes for elaborate distributions. *Bayesian Analysis* **6**, Number 4, 1–48.

WELHAM, S.J., CULLIS, B.R., KENWARD, M.G. & THOMPSON, R. (2007). A comparison of mixed model splines for curve fitting. *Aust. N. Z. J. Stat.* **49**, 1–23.

WILLIS, J.K., ROEMMICH, D. & CORNUELLE, B. (2004). Interannual variability in upper ocean heat content, temperature, and thermosteric expansion on global scales. *J. Geophys. Res.* **109**, C12036, 1–13.

WOOD, S.N. (2003). Thin-plate regression splines. *J. R. Statist. Soc. Ser. B* **65**, 95–114.

YEE, T.W. & STEPHENSON, A.G. (2007). Vector generalized linear and additive extreme value models. *Extremes* **10**, 1–19.