Supplementary material for "Max-and-Smooth: a two-step approach for approximate Bayesian inference in latent Gaussian models"

Birgir Hrafnkelsson[*,¶], Stefan Siegert[†], Raphaël Huser[‡], Haakon Bakka[§] and Árni V. Jóhannesson[*]

1 An approximate inference scheme for LGMs with a multivariate link function

In this section, we show the details of Max-and-Smooth, in particular, Step 2 which involves smoothing the estimates of the latent parameters jointly. Max-and-Smooth is used to infer the parameters of the extend LGM presented in Section 3.2 of the main paper, and it is based on the approximations to the posterior density in equations (3.1) and (3.2) in the main paper. We consider a model that is such that either $\hat{\boldsymbol{\eta}}$ or $\tilde{\boldsymbol{\eta}}$ is treated as the data. The proposed data density is either $\mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}^{-1})$ where $Q_{\eta y}$ is known and $Q_{\eta y} = \Sigma_{\eta y}^{-1}$, or $\mathrm{N}(\tilde{\boldsymbol{\eta}}|\boldsymbol{\eta}, \tilde{Q}_{\eta y}^{-1})$ where $\tilde{Q}_{\eta y}$ is known and $\tilde{Q}_{\eta y} = \Omega_{\eta y}^{-1}$. The numerical values of the matrices, $\Sigma_{\eta y}$ and $\Omega_{\eta y}$ are evaluated from the already observed data and these two matrices stem from the matrices $\Sigma_{\eta y_i}$ and $\Omega_{\eta y_i}$, $i = 1, ..., G$, respectively, which are defined in Section 3.3 in the main paper. The model for $\hat{\boldsymbol{\eta}}$ can be written hierarchically as

$$\begin{aligned}\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}, \boldsymbol{\theta}) &= \mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}^{-1}),\\ \pi(\boldsymbol{\eta}|\boldsymbol{\nu}, \boldsymbol{\theta}) &= \mathrm{N}(\boldsymbol{\eta}|Z\boldsymbol{\nu}, Q_{\epsilon}^{-1}),\\ \pi(\boldsymbol{\nu}|\boldsymbol{\theta}) &= \mathrm{N}(\boldsymbol{\nu}|\boldsymbol{\mu}_{\nu}, Q_{\nu}^{-1})\end{aligned}$$

and $\pi(\boldsymbol{\theta})$ is the prior density for $\boldsymbol{\theta}$ as before. The model for $\tilde{\boldsymbol{\eta}}$ has the same structure, however, $\hat{\boldsymbol{\eta}}$ and $Q_{\eta y}$ are replaced with $\tilde{\boldsymbol{\eta}}$ and $\tilde{Q}_{\eta y}$, respectively. The posterior density for the model above is given by

$$\begin{aligned}\pi(\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\theta}|\hat{\boldsymbol{\eta}}) &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}, \boldsymbol{\theta})\\ &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{\theta})\mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}^{-1})\\ &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{\theta})\mathrm{N}(\boldsymbol{\eta}|\hat{\boldsymbol{\eta}}, Q_{\eta y}^{-1})\\ &\propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\eta}, \boldsymbol{\nu}|\boldsymbol{\theta})\hat{L}(\boldsymbol{\eta}|\boldsymbol{y}).\end{aligned}$$

The above posterior density stems from looking at it as a function of $\boldsymbol{\eta}$ and taking $\hat{\boldsymbol{\eta}}$ as a fixed quantity, which gives $\mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}^{-1}) = \mathrm{N}(\boldsymbol{\eta}|\hat{\boldsymbol{\eta}}, Q_{\eta y}^{-1})$. So, the above posterior density

[*]University of Iceland, Reykjavik, Iceland, birgirhr@hi.is, avj2@hi.is
[†]University of Exeter, Exeter, UK, s.siegert@exeter.ac.uk
[‡]King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, raphael.huser@kaust.edu.sa
[§]University of Oslo, Oslo, Norway, bakka@r-inla.org
[¶]Corresponding author

is exactly the same as the approximated posterior density in equation (3.1) in the main paper, and $\hat{L}(\boldsymbol{\eta}|\boldsymbol{y})$ is the Gaussian approximation of the original likelihood function, $L(\boldsymbol{\eta}|\boldsymbol{y})$, scaled with a constant. The model presented in this section is a Gaussian–Gaussian model and it is more convenient to approach the inference for the unknown parameters through this model than the original model with the approximated likelihood function even though the two posterior densities are the same. In particular, it will be useful to assume that $\hat{\boldsymbol{\eta}}$ is Gaussian when deriving the marginal posterior density of $\boldsymbol{\theta}$ and the conditional posterior density of $(\boldsymbol{\eta}, \boldsymbol{\nu})$ given $\boldsymbol{\theta}$. To tackle the inference for $(\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\theta})$ the joint prior density of $\boldsymbol{x} = (\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\nu}^\mathsf{T})^\mathsf{T}$ conditional on $\boldsymbol{\theta}$ is derived. It is given by

$$\pi(\boldsymbol{x}|\boldsymbol{\theta}) = \mathrm{N}\left(\boldsymbol{x} \middle| \begin{bmatrix} Z\boldsymbol{\mu}_\nu \\ \boldsymbol{\mu}_\nu \end{bmatrix}, \begin{bmatrix} Q_\epsilon & -Q_\epsilon Z \\ -Z^\mathsf{T} Q_\epsilon & Q_\nu + Z^\mathsf{T} Q_\epsilon Z \end{bmatrix}^{-1}\right).$$

The covariance matrix of $\boldsymbol{x}$ given $\boldsymbol{\theta}$ is

$$\Sigma_x = Q_x^{-1} = \begin{bmatrix} Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T} & ZQ_\nu^{-1} \\ Q_\nu^{-1}Z^\mathsf{T} & Q_\nu^{-1} \end{bmatrix}.$$

Let $p = \dim(\boldsymbol{\eta}) = MG$, $q = \dim(\boldsymbol{\nu})$ and $B = \begin{bmatrix} I_{p\times p} & 0_{p\times q} \end{bmatrix}$. Because $\hat{\boldsymbol{\eta}}$ can be expressed as $\hat{\boldsymbol{\eta}} = \boldsymbol{\eta} + \boldsymbol{e} = B\boldsymbol{x} + \boldsymbol{e}$, where $\boldsymbol{e} \sim \mathrm{N}(\boldsymbol{0}, Q_{\eta y}^{-1})$, then the joint density of $(\hat{\boldsymbol{\eta}}^\mathsf{T}, \boldsymbol{x}^\mathsf{T})^\mathsf{T} = (\hat{\boldsymbol{\eta}}^\mathsf{T}, \boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\nu}^\mathsf{T})^\mathsf{T}$ is given by

$$\pi\left(\begin{bmatrix} \hat{\boldsymbol{\eta}} \\ \boldsymbol{x} \end{bmatrix}\right) = \mathrm{N}\left(\begin{bmatrix} \hat{\boldsymbol{\eta}} \\ \boldsymbol{x} \end{bmatrix} \middle| \begin{bmatrix} B\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_x \end{bmatrix}, \begin{bmatrix} Q_{\eta y}^{-1} + BQ_x^{-1}B^\mathsf{T} & BQ_x^{-1} \\ Q_x^{-1}B^\mathsf{T} & Q_x^{-1} \end{bmatrix}\right).$$

The mean vector and the covariance matrix of $(\hat{\boldsymbol{\eta}}^\mathsf{T}, \boldsymbol{x}^\mathsf{T})^\mathsf{T}$ can, respectively, be written as

$$\begin{bmatrix} Z\boldsymbol{\mu}_\nu \\ Z\boldsymbol{\mu}_\nu \\ \boldsymbol{\mu}_\nu \end{bmatrix}, \quad \begin{bmatrix} Q_{\eta y}^{-1} + Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T} & Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T} & ZQ_\nu^{-1} \\ Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T} & Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T} & ZQ_\nu^{-1} \\ Q_\nu^{-1}Z^\mathsf{T} & Q_\nu^{-1}Z^\mathsf{T} & Q_\nu^{-1} \end{bmatrix}.$$

The precision matrix of $(\hat{\boldsymbol{\eta}}^\mathsf{T}, \boldsymbol{x}^\mathsf{T})^\mathsf{T}$ is

$$Q_{\hat{\eta}x} = \Sigma_{\hat{\eta}x}^{-1} = \begin{bmatrix} Q_{\eta y} & -Q_{\eta y}B \\ -B^\mathsf{T}Q_{\eta y} & Q_x + B^\mathsf{T}Q_{\eta y}B \end{bmatrix}$$

and it can be written as

$$Q_{\hat{\eta}x} = \begin{bmatrix} Q_{\eta y} & -Q_{\eta y} & 0 \\ -Q_{\eta y} & Q_\epsilon + Q_{\eta y} & -Q_\epsilon Z \\ 0 & -Z^\mathsf{T}Q_\epsilon & Q_\nu + Z^\mathsf{T}Q_\epsilon Z \end{bmatrix}.$$

The marginal density of $\hat{\boldsymbol{\eta}}$ given $\boldsymbol{\theta}$ is

$$\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta}) = \mathrm{N}(\hat{\boldsymbol{\eta}}|B\boldsymbol{\mu}_x, Q_{\eta y}^{-1} + BQ_x^{-1}B^\mathsf{T}) = \mathrm{N}(\hat{\boldsymbol{\eta}}|Z\boldsymbol{\mu}_\nu, Q_{\eta y}^{-1} + Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^\mathsf{T}).$$

The marginal posterior density of $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\eta}}$ is $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}}) \propto \pi(\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta})$ and it can be represented as

$$\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}}) \propto \pi(\boldsymbol{\theta}) \frac{\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}}, \boldsymbol{\theta})}, \tag{1.1}$$

using the fact that $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta})\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\eta}}|\boldsymbol{x},\boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})$. The densities $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{x},\boldsymbol{\theta})$ and $\pi(\boldsymbol{x}|\boldsymbol{\theta})$ have precision matrices $Q_{\eta y}$ and Q_x, respectively. The precision matrix of $\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta})$ is given below. Note that the density $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}})$ can be evaluated with any value for $\boldsymbol{x}$ because it does not depend on $\boldsymbol{x}$. This fact can be used to simplify calculations of the marginal density in (1.1) by setting $\boldsymbol{x}$ equal to a convenient value, e.g., $\boldsymbol{x} = \boldsymbol{0}$. The conditional posterior density of $\boldsymbol{x}$ given $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\theta}$ is

$$\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{x}|Q_{x|\hat{\eta}}^{-1}(Q_x\boldsymbol{\mu}_x + B^{\mathsf{T}}Q_{\eta y}\hat{\boldsymbol{\eta}}), Q_{x|\hat{\eta}}^{-1}), \tag{1.2}$$

where

$$Q_{x|\hat{\eta}} = Q_x + B^{\mathsf{T}}Q_{\eta y}B = \begin{bmatrix} Q_\epsilon + Q_{\eta y} & -Q_\epsilon Z \\ -Z^{\mathrm{T}}Q_\epsilon & Q_\nu + Z^{\mathsf{T}}Q_\epsilon Z \end{bmatrix} \tag{1.3}$$

and

$$Q_x\boldsymbol{\mu}_x + B^{\mathsf{T}}Q_{\eta y}\hat{\boldsymbol{\eta}} = \begin{bmatrix} Q_{\eta y}\hat{\boldsymbol{\eta}} \\ Q_\nu\boldsymbol{\mu}_\nu \end{bmatrix}. \tag{1.4}$$

To sample from the posterior density of $(\boldsymbol{x},\boldsymbol{\theta})$, first, a sample from $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}})$, which is an approximation of $\pi(\boldsymbol{\theta}|\boldsymbol{y})$, is taken. Then a sample from $\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta})$, which is an approximation of $\pi(\boldsymbol{x}|\boldsymbol{y},\boldsymbol{\theta})$, is taken.

In order to gain insight into how $\hat{\boldsymbol{\eta}}$ is used to update our prior knowledge of $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ then the conditional distribution of $\boldsymbol{\eta}$ given $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\theta}$ is explored as well as the conditional distribution of $\boldsymbol{\nu}$ given $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\theta}$. The density of the conditional distribution of $\boldsymbol{\eta}$ can be found by using the fact that

$$\pi(\boldsymbol{\eta}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta}) \propto \pi(\boldsymbol{\eta}|\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta})$$

where $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}) = \mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q_{\eta y}^{-1})$ and $\pi(\boldsymbol{\eta}|\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\eta}|Z\boldsymbol{\mu}_\nu, Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^{\mathrm{T}})$. The result is a Gaussian density with mean $\boldsymbol{\mu}_{\eta|\hat{\eta}}$ and precision matrix $Q_{\eta|\hat{\eta}}$, where

$$\boldsymbol{\mu}_{\eta|\hat{\eta}} = Q_{\eta|\hat{\eta}}^{-1}\{(Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^{\mathsf{T}})^{-1}Z\boldsymbol{\mu}_\nu + Q_{\eta y}\hat{\boldsymbol{\eta}}\}, \quad Q_{\eta|\hat{\eta}} = (Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^{\mathsf{T}})^{-1} + Q_{\eta y}.$$

It is clear from the above formula for the posterior mean $\boldsymbol{\mu}_{\eta|\hat{\eta}}$, that it is a linear combination of the prior mean $Z\boldsymbol{\mu}_\nu$ and the maximum likelihood estimates in $\hat{\boldsymbol{\eta}}$. The posterior mean stems from Bayesian regression where the design matrix is an identity matrix, the error precision matrix is $Q_{\eta y}$, and the prior precision matrix is $Q_\eta = (Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^{\mathsf{T}})^{-1}$. The prior precision matrix assigns weights to the prior mean $Z\boldsymbol{\mu}_\nu$ while the precision matrix $Q_{\eta y}$ assigns weights to $\hat{\boldsymbol{\eta}}$. The larger the weights in $Q_{\eta y}$ with respect to Q_η, the greater is the influence of $\hat{\boldsymbol{\eta}}$ on the posterior mean of $\boldsymbol{\eta}$. If the prior mean of $\boldsymbol{\mu}_\nu$ is equal to zero then the posterior mean of $\boldsymbol{\eta}$ is simply

$$\boldsymbol{\mu}_{\eta|\hat{\eta}} = Q_{\eta|\hat{\eta}}^{-1}Q_{\eta y}\hat{\boldsymbol{\eta}} = \{Q_{\eta y}^{-1}(Q_\epsilon^{-1} + ZQ_\nu^{-1}Z^{\mathsf{T}})^{-1} + I\}^{-1}\hat{\boldsymbol{\eta}} = (Q_{\eta y}^{-1}Q_\eta + I)^{-1}\hat{\boldsymbol{\eta}},$$

so, if the elements of $Q_{\eta y}$ are large with respect to the elements of Q_η, $\boldsymbol{\mu}_{\eta|\hat{\eta}}$ will be closer to $\hat{\boldsymbol{\eta}}$.

The conditional density of $\boldsymbol{\nu}$ given $\hat{\boldsymbol{\eta}}$ can be found by using the fact that

$$\pi(\boldsymbol{\nu}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta}) \propto \pi(\boldsymbol{\nu}|\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\nu},\boldsymbol{\theta}),$$

where $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\nu},\boldsymbol{\theta}) = \mathrm{N}(\hat{\boldsymbol{\eta}}|Z\boldsymbol{\nu}, Q_{\eta y}^{-1} + Q_{\epsilon}^{-1})$ and $\pi(\boldsymbol{\nu}|\boldsymbol{\theta}) = \mathrm{N}(\boldsymbol{\nu}|\boldsymbol{\mu}_{\nu}, Q_{\nu}^{-1})$. The result is a Gaussian density with mean $\boldsymbol{\mu}_{\nu|\hat{\eta}}$ and precision matrix $Q_{\nu|\hat{\eta}}$ where

$$\boldsymbol{\mu}_{\nu|\hat{\eta}} = Q_{\nu|\hat{\eta}}^{-1}\{Q_{\nu}\boldsymbol{\mu}_{\nu} + Z^{\mathsf{T}}(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}\hat{\boldsymbol{\eta}}\}, \quad Q_{\nu|\hat{\eta}} = Q_{\nu} + Z^{\mathsf{T}}(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}Z.$$

The formula for $\boldsymbol{\mu}_{\nu|\hat{\eta}}$ shows that the posterior mean for $\boldsymbol{\nu}$ is a linear combination of the prior mean $\boldsymbol{\mu}_{\nu}$ and the maximum likelihood estimates in $\hat{\boldsymbol{\eta}}$. This posterior mean is the result of a Bayesian regression with design matrix Z, the error precision matrix $(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}$ and the prior precision matrix Q_{ν}. The larger the weights in $(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}$ are with respect to Q_{ν}, the greater is the influence of $\hat{\boldsymbol{\eta}}$ on the posterior mean $\boldsymbol{\mu}_{\nu|\hat{\eta}}$. If the prior mean of $\boldsymbol{\mu}_{\nu}$ is equal to zero then the posterior mean of $\boldsymbol{\nu}$ is

$$\boldsymbol{\mu}_{\nu|\hat{\eta}} = \{Q_{\nu} + Z^{\mathsf{T}}(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}Z\}^{-1}Z^{\mathsf{T}}(Q_{\epsilon}^{-1} + Q_{\eta y}^{-1})^{-1}\hat{\boldsymbol{\eta}},$$

thus, in this case $\boldsymbol{\mu}_{\nu|\hat{\eta}}$ is solely a linear combination of $\hat{\boldsymbol{\eta}}$, and the role of the prior precision matrix Q_{ν} in $\boldsymbol{\mu}_{\nu|\hat{\eta}}$ can be seen.

The marginal posterior density of $\boldsymbol{\theta}$ given $\hat{\boldsymbol{\eta}}$ is proportional to $\pi(\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta})$ and it can be represented in terms of $\boldsymbol{\nu}$, that is,

$$\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}}) \propto \pi(\boldsymbol{\theta})\frac{\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\nu},\boldsymbol{\theta})\pi(\boldsymbol{\nu}|\boldsymbol{\theta})}{\pi(\boldsymbol{\nu}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta})}, \tag{1.5}$$

using the fact that $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta})\pi(\boldsymbol{\nu}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta}) = \pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\nu},\boldsymbol{\theta})\pi(\boldsymbol{\nu}|\boldsymbol{\theta})$, where $\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\nu},\boldsymbol{\theta})$, $\pi(\boldsymbol{\nu}|\boldsymbol{\theta})$ and $\pi(\boldsymbol{\nu}|\hat{\boldsymbol{\eta}},\boldsymbol{\theta})$ are as above. Representing $\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}})$ in terms of $\boldsymbol{\nu}$ as in (1.5) is useful for LGMs with a univariate link function. Furthermore, this setup is relevant to LGMs with a multivariate link function when the approximated likelihood induces independence in such a way that one or more of the $\boldsymbol{\eta}_m$ vectors in $\boldsymbol{\eta}$ can be inferred independently of the other vectors in $\boldsymbol{\eta}$.

2 Examples: simulated data

2.1 Log-variance on a lattice

This example is presented in the main paper. In this section the details of the statistical model and its inference are given. Results are given in Section 4 of the main paper.

Statistical model

Let $y_{i,t}$ be the observed variable at lattice point i and time t, respectively. The index i corresponds to the lattice point with horizontal coordinates i_1 and vertical coordinates i_2. We assume that the observation at lattice point i and time point t follows a Gaussian distribution with mean zero and variance σ_i^2, i.e.,

$$y_{i,t} \sim \mathrm{N}(0, \sigma_i^2), \quad \sigma_i^2 > 0,$$

where $i \in \{1, ..., N\}$, $t \in \{1, ..., T\}$, N is the number of lattice points and T is the number of observations at each lattice point. Here the "groups" mentioned in Section

2.2 in the main paper represent the observed variables grouped according to the lattice points, thus, the number of groups is $G = N$. The data density for $\boldsymbol{y}_i = (y_{i,1}, ..., y_{i,T})^\mathsf{T}$ is then

$$\pi(\boldsymbol{y}_i|\sigma_i^2) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{y_{i,t}^2}{2\sigma_i^2} \right). \tag{2.1}$$

The parameter σ_i^2 is transformed to $x_i = \log(\sigma_i^2)$, so $x_i \in \mathbb{R}$. In terms of x_i, this density can be written as

$$\pi(\boldsymbol{y}_i|x_i) = (2\pi)^{-T/2} \exp(-(T/2)x_i) \exp\left(-\frac{\exp(-x_i)}{2} \sum_{t=1}^{T} y_{i,t}^2 \right). \tag{2.2}$$

The parameters are collected into the vector $\boldsymbol{x} = (x_1, ..., x_N)^\mathsf{T}$, and modeled at the latent level as a Gaussian Markov random field (GMRFs) on a regular lattice (Rue and Held, 2005). The precision matrix of $\boldsymbol{x}$ is $Q_x = \tau Q_u$ where

$$Q_u = R_{n_1} \otimes I_{n_2} + I_{n_1} \otimes R_{n_2}, \tag{2.3}$$

τ is a precision parameter, $\otimes$ denotes the Kronecker product, n_1 and n_2 are the dimensions of the lattice ($N = n_1 n_2$), I_m is an m dimensional identity matrix, and

$$R_m = \begin{bmatrix} 2 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \vdots & \vdots & \vdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 2 \end{bmatrix} \tag{2.4}$$

is an $m \times m$ matrix. Note that R_m is different from the R matrix defined in Rue and Held (2005, Section 3.3.1, p. 95), in (2.4) the first and the last diagonal elements are equal to 2 as opposed to 1 and, as a result, the precision matrix in (2.3) is full rank as opposed to having rank $N - 1$ as the Q matrix defined in Rue and Held (2005, Section 3.3.2, p. 107). The precision matrix in (2.3) corresponds to a random process defined on a lattice such that the variables on the points surrounding it are all equal to zero.

Here τ is set as the hyperparameter. At the hyperparameter level we assign a gamma prior density to τ, such that

$$\tau \sim \text{Gamma}(10, 10).$$

In our simulation study based on this model, we set the value of the hyperparameter equal to $\tau = 1.0$, and the models were inferred with the unstructured model term set equal to zero, i.e., $\boldsymbol{\epsilon} = \mathbf{0}$. Furthermore, the size of the lattice was set equal to 10×10, 20×20, 35×35, and 50×50, so $N = 100$, $N = 400$, $N = 1225$ and $N = 2500$, and the number of observations per lattice point is set equal to $T = 10, 20, 50, 100$.

Inference

We use Max-and-Smooth to infer $\boldsymbol{x}$ and τ. The likelihood function for x_i was approximated with two Gaussian approximations. The first one has mean equal to the mode of the likelihood function, namely

$$\hat{x}_i = \log\left(\frac{1}{T}\sum_{t=1}^{T} y_{i,t}^2\right)$$

and variance equal to the inverse of the observed information,

$$\mathcal{I}_{xy_i}^{-1} = \frac{2}{T}.$$

The likelihood function was also approximated with a Gaussian density with mean and variance equal to those of the normalized likelihood function. Here the normalized likelihood function is a log-inverse-gamma density, which has the general form

$$\pi(x_i) = \frac{\gamma^\alpha}{\Gamma(\alpha)}\exp\{-\alpha x_i - \gamma\exp(-x_i)\}$$

and its mean and variance are

$$\mathrm{E}(x_i) = -\psi(\alpha) + \log(\gamma), \quad \mathrm{var}(x_i) = \psi'(\alpha)$$

where $\psi(\cdot)$ and $\psi'(\cdot)$ are the digamma and trigamma functions, respectively. The normalized likelihood function of x_i is a log-inverse-gamma density with parameters $\alpha = T/2$ and $\gamma = 0.5\sum_{t=1}^{T} y_{i,t}^2$, and its mean is

$$\mathrm{E}(x_i|\boldsymbol{y}_i) = -\psi(T/2) + \log\left(0.5\sum_{t=1}^{T} y_{i,t}^2\right) = \hat{x}_i + \log(T/2) - \psi(T/2),$$

and its variance is $\mathrm{var}(x_i|\boldsymbol{y}_i) = \psi'(T/2)$. Denote the above mean by $\tilde{x}_i$. Thus, the likelihood function was approximated with both $L(x_i) \approx c_i\hat{L}(x_i) = c_i\mathrm{N}(x_i|\hat{x}_i, 2/T)$, and $L(x_i) \approx \tilde{c}_i\tilde{L}(x_i) = \tilde{c}_i\mathrm{N}(x_i|\tilde{x}_i, \psi'(T/2))$, where c_i and $\tilde{c}_i$ are constants independent of x_i.

The inference for the above model involves modeling the data in $\hat{\boldsymbol{x}} = (\hat{x}_1, ..., \hat{x}_N)^\mathsf{T}$ (or in $\tilde{\boldsymbol{x}} = (\tilde{x}_1, ..., \tilde{x}_N)^\mathsf{T}$) with a pseudo LGM with a univariate link function. This model is a Gaussian–Gaussian model, and it can be inferred by using a slightly modified version of the inference scheme in Section 1 and Section 3.3 in the main paper. The data density for $\hat{x}_i$ in the pseudo model is given by

$$\pi(\hat{x}_i|x_i) = \mathrm{N}(\hat{x}_i|x_i, 2/T) = \frac{1}{\sqrt{2\pi(2/T)}}\exp\left\{-\frac{(\hat{x}_i - x_i)^2}{2(2/T)}\right\}, \tag{2.5}$$

and the latent level and the hyperparameter level of the pseudo model are the same as the ones in the original model. The data density for $\tilde{x}_i$ in the pseudo model is $\pi(\tilde{x}_i|x_i) = \mathrm{N}(\tilde{x}_i|x_i, \psi'(T/2))$.

The inference scheme below is based on the pseudo model and is presented in terms of $\hat{\boldsymbol{x}}$ as opposed to $\tilde{\boldsymbol{x}}$. However, the pseudo model below can be modified to be applicable for $\tilde{\boldsymbol{x}}$, in particular, $\hat{\boldsymbol{x}}$ is replaced with $\tilde{\boldsymbol{x}}$ and the variance, $2/T$, is replaced with the variance $\psi'(T/2)$. The likelihood of $\boldsymbol{x}$, $L(\boldsymbol{x})$, is approximated with $\hat{L}(\boldsymbol{x})$, that is

$$L(\boldsymbol{x}) = \prod_{i=1}^{N} L(x_i) \approx c\hat{L}(\boldsymbol{x}) = \prod_{i=1}^{N} c_i \hat{L}(x_h) = c\mathrm{N}(\boldsymbol{x}|\hat{\boldsymbol{x}}, (2/T)I)$$

where $c = \prod_{i=1}^{N} c_i$. Here $\hat{\boldsymbol{x}}$ is taken as the data and the covariance matrix $(2/T)I$ is fixed. The approximate data density is then

$$\pi(\hat{\boldsymbol{x}}|\boldsymbol{x}) = \mathrm{N}(\hat{\boldsymbol{x}}|\boldsymbol{x}, (2/T)I).$$

Here the hyperparameter is $\theta = \tau$. To sample from the posterior density of $(\boldsymbol{x}, \theta)$, we first draw a sample from the marginal posterior density of θ and then we sample from the posterior density of $\boldsymbol{x}$ conditional on $\hat{\boldsymbol{x}}$ and θ. When equation (1.5) is applied to the model with $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{x}}$ and $\boldsymbol{\nu} = \boldsymbol{x}$ then

$$\pi(\theta|\hat{\boldsymbol{x}}) \propto \pi(\theta)\pi(\hat{\boldsymbol{x}}|\theta) = \pi(\theta)\frac{\pi(\hat{\boldsymbol{x}}|\boldsymbol{x}, \theta)\pi(\boldsymbol{x}|\theta)}{\pi(\boldsymbol{x}|\hat{\boldsymbol{x}}, \theta)},$$

where

$$\pi(\hat{\boldsymbol{x}}|\boldsymbol{x}, \theta) = \mathrm{N}(\hat{\boldsymbol{x}}|\boldsymbol{x}, (2/T)I),$$

$$\pi(\boldsymbol{x}|\hat{\boldsymbol{x}}, \theta) = \mathrm{N}(\boldsymbol{x}|\boldsymbol{\mu}_{x|\hat{x}}, \Sigma_{x|\hat{x}}),$$

$$\Sigma_{x|\hat{x}} = \{Q_x + (T/2)I\}^{-1} = \{\theta Q_u + (T/2)I\}^{-1},$$

$$\hat{\boldsymbol{\mu}}_{x|\hat{x}} = \Sigma_{x|\hat{x}}(T/2)\hat{\boldsymbol{x}} = \{(2/T)\theta Q_u + I\}^{-1}\hat{\boldsymbol{x}},$$

and the prior density for $\boldsymbol{x}$, $\pi(\boldsymbol{x}|\theta)$, is Gaussian with mean zero and precision matrix $Q_x = \theta Q_u$ where Q_u is specified with equation (2.3). Samples from the posterior density of $\boldsymbol{x}$ conditional on $\hat{\boldsymbol{x}}$ and θ are obtained from the Gaussian density $\pi(\boldsymbol{x}|\hat{\boldsymbol{x}}, \theta) = \mathrm{N}(\boldsymbol{x}|\boldsymbol{\mu}_{x|\hat{x}}, \Sigma_{x|\hat{x}})$ presented above.

2.2 Linear regression on a lattice

Statistical model

Let $y_{i,t}$ and $f_{i,t}$ be the response and the predictor at lattice point i and time t, respectively. The data are assumed to follow a Gaussian distribution with mean

$$\alpha_i + \beta_i(f_{i,t} - \bar{f}_i),$$

where $\bar{f}_i = T^{-1}\sum_{t=1}^{T} f_{i,t}$, and variance σ_i^2, i.e.,

$$y_{i,t} \sim \mathrm{N}(\alpha_i + \beta_i(f_{i,t} - \bar{f}_i), \sigma_i^2), \quad \alpha_i, \beta_i \in \mathbb{R}, \quad \sigma_i^2 > 0,$$

where $i \in \{1, ..., N\}$, $t \in \{1, ..., T\}$, N is the number of lattice points and T is the number of observations at each lattice point. The lattice points form the groups, so the number of groups is $G = N$. The data density for $\boldsymbol{y}_i = (y_{i1}, ..., y_{iT})^\mathsf{T}$ is then

$$\pi(\boldsymbol{y}_i|\alpha_i, \beta_i, \sigma_i^2) = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\{y_{i,t} - \alpha_i - \beta_i(f_{i,t} - \bar{f}_i)\}^2}{2\sigma_i^2} \right]. \tag{2.6}$$

The parameter σ_i^2 is transformed to $\tau_i = \log(\sigma_i^2)$, so $\tau_i \in \mathbb{R}$.

The parameters are collected into the vectors $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)^\mathsf{T}$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_N)^\mathsf{T}$ and $\boldsymbol{\tau} = (\tau_1, ..., \tau_N)^\mathsf{T}$. The latent level consists of three linear predictors for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ of the form

$$\begin{aligned} \boldsymbol{\alpha} &= \boldsymbol{u}_\alpha + \boldsymbol{\epsilon}_\alpha, \\ \boldsymbol{\beta} &= \boldsymbol{u}_\beta + \boldsymbol{\epsilon}_\beta, \\ \boldsymbol{\tau} &= \boldsymbol{u}_\tau + \boldsymbol{\epsilon}_\tau, \end{aligned} \tag{2.7}$$

where $\boldsymbol{u}_\alpha$, $\boldsymbol{u}_\beta$ and $\boldsymbol{u}_\tau$ are modeled as first-order intrinsic Gaussian Markov random fields (IGMRFs) on a regular lattice (Rue and Held, 2005, Section 3.3.2, pp. 104–108). Selecting Gaussian Markov random fields as priors for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ is justified as spatial dependancy between elements close in space is expected. This model choice is beneficial for the estimation of these parameters as it allows borrowing strength from neighboring lattice points.

A priori the vectors $\boldsymbol{u}_\alpha$, $\boldsymbol{u}_\beta$ and $\boldsymbol{u}_\tau$ are assumed to be independent. Their precision matrices are $Q_{u,\alpha}$, $Q_{u,\beta}$ and $Q_{u,\tau}$ where $Q_{u,l} = \sigma_{u,l}^{-2} Q_u$, $l \in \{\alpha, \beta, \tau\}$, and Q_u is the precision matrix given by equation (2.8) below, and $\sigma_{u,\alpha}$, $\sigma_{u,\beta}$ and $\sigma_{u,\tau}$ are unknown hyperparameters. The dimensions of the lattice are n_1 and n_2 and $N = n_1 n_2$. The form of Q_u is

$$Q_u = R_{n_1} \otimes I_{n_2} + I_{n_1} \otimes R_{n_2} \tag{2.8}$$

where $\otimes$ denotes the Kronecker product, I_m is an m dimensional identity matrix and

$$R_m = \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \vdots & \vdots & \vdots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 2 & -1 \\ & & & & & -1 & 1 \end{bmatrix} \tag{2.9}$$

is an $m \times m$ matrix (Rue and Held, 2005, Section 3.3.1, p. 95). Let u_{l,i_1,i_2} be the element of $\boldsymbol{u}_l$ with horizontal coordinates i_1 and vertical coordinates i_2. If (i_1, i_2) is an interior lattice point then the conditional density of u_{l,i_1,i_2} conditional on $\boldsymbol{u}_{l,-(i_1,i_2)}$ (the elements of $\boldsymbol{u}_l$ other than u_{l,i_1,i_2}) is Gaussian with mean

$$\frac{1}{4}(u_{l,i_1+1,i_2} + u_{l,i_1-1,i_2} + u_{l,i_1,i_2+1} + u_{l,i_1,i_2-1})$$

and variance $0.25\sigma^2_{u,l}$. The precision matrix Q_u is not full rank, its rank is $N-1$. By conditioning on $\sum_i u_{l,i} = \mathbf{1}^\mathsf{T}\boldsymbol{u}_l = u^*_l$ then the density for $\boldsymbol{u}_l$ can be specified with the proper density $\pi(\boldsymbol{u}_l) = \pi(\boldsymbol{u}_l|u^*_l)\pi(u^*_l)$ where $\pi(u^*_l)$ is a Gaussian density with mean zero and precision γ, and

$$\log \pi(\boldsymbol{u}_l|u^*_l) = -\frac{(N-1)}{2}\log(2\pi) - \frac{(N-1)}{2}\log(\sigma^2_{u,l}) + \frac{1}{2}\sum_{i=2}^{N}\log(\lambda_i) - \frac{\sigma^{-2}_{u,l}}{2}\boldsymbol{u}_l^\mathsf{T} Q_u \boldsymbol{u}_l, \tag{2.10}$$

where λ_i is the i-th eigenvalue of Q_u,

$$\lambda_i = 2(1-\cos(\pi(i_1-1)/n_1)) + 2(1-\cos(\pi(i_2-1)/n_2)),$$

$i_1 = 1,...,n_1$, $i_2 = 1,...,n_2$, $i = i_1 + n_1(i_2-1)$. Note that the first eigenvalue ($i=1$) is excluded in the above density. By letting $\gamma \to 0$ then $\pi(\boldsymbol{u}_l) \propto \pi(\boldsymbol{u}_l|u^*_l)$, and $\pi(\boldsymbol{u}_l|u^*_l)$ can be used to infer $\boldsymbol{u}_l$ and $\sigma_{u,l}$.

The three vectors $\boldsymbol{\epsilon}_\alpha$, $\boldsymbol{\epsilon}_\beta$ and $\boldsymbol{\epsilon}_\tau$ contain unstructured model errors with variances $\sigma^2_{\epsilon,\alpha}, \sigma^2_{\epsilon,\beta}$ and $\sigma^2_{\epsilon,\tau}$, respectively. The prior distributions for $\boldsymbol{\epsilon}_\alpha$, $\boldsymbol{\epsilon}_\beta$ and $\boldsymbol{\epsilon}_\tau$ can be written as

$$\boldsymbol{\epsilon}_\alpha \sim \mathrm{N}(0, \sigma^2_{\epsilon,\alpha}\mathrm{I}), \quad \boldsymbol{\epsilon}_\beta \sim \mathrm{N}(0, \sigma^2_{\epsilon,\beta}\mathrm{I}), \quad \boldsymbol{\epsilon}_\tau \sim \mathrm{N}(0, \sigma^2_{\epsilon,\tau}\mathrm{I}).$$

At the hyperparameter level an exponential prior density is assigned to each of the parameters $\sigma_{u,\alpha}$, $\sigma_{\epsilon,\alpha}$, $\sigma_{u,\beta}$, $\sigma_{\epsilon,\beta}$, $\sigma_{u,\tau}$ and $\sigma_{\epsilon,\tau}$. The decision to select independent exponential distributions for these parameters is motivated by the penalized complexity (PC) prior framework presented in Simpson et al. (2017). The PC prior framework can be used to design prior densities for hyperparameters like the standard deviation parameters of random effects. It is based on the Kullback-Leibler divergence and is designed to support simpler models. One of the principles of this framework is such that in the case of standard deviation parameters their PC prior densities support the case of them being equal to zero. If the data suggest that a given parameter is equal to zero then the PC prior density pulls the posterior mass toward zero. On the other hand, evidence for a more complex model with a standard deviation parameter different from zero needs to be pulled by the data.

Inference

We apply Max-and-Smooth to infer the unknown parameters of the extended LGM above, see Section 1 and Section 3.3 in the main paper. The extended LGM model has a trivariate link function. Let $\boldsymbol{\psi}_i = (\alpha_i, \beta_i, \sigma^2_i)^\mathsf{T}$ and let $g : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}^3$ be a trivariate link function such that $g(x_1, x_2, x_3) = (x_1, x_2, \log(x_3))^\mathsf{T}$, and $g^{-1}(u_1, u_2, u_3) = (u_1, u_2, \exp(u_3))^\mathsf{T}$. Let

$$\boldsymbol{\eta}_i := (\alpha_i, \beta_i, \tau_i)^\mathsf{T} = g(\boldsymbol{\psi_i}) = (\alpha_i, \beta_i, \log(\sigma^2_i))^\mathsf{T}.$$

Now define $\hat{\boldsymbol{\psi}}_i := (\hat{\alpha}_i, \hat{\beta}_i, \hat{\sigma}^2_i)^\mathsf{T}$ as the maximum likelihood estimate of $\boldsymbol{\psi}_i$ based on the data density for lattice point i given by (2.6) which is computed separately from the

other lattice points. Let F_i be the design matrix with $(1, f_{i,t} - \bar{f}_i)$ as its t-th row. Then $(\hat{\alpha}_i, \hat{\beta}_i)^\mathsf{T} = (F_i^\mathsf{T} F_i)^{-1} F_i^\mathsf{T} \boldsymbol{y}_i$ and $\hat{\sigma}_i^2 = T^{-1} \sum_{t=1}^{T} \{y_{i,t} - \hat{\alpha}_i - \hat{\beta}_i (f_{i,t} - \bar{f}_i)\}^2$.

Instead of using $\boldsymbol{y}_i$ as the data, we treat

$$\hat{\boldsymbol{\eta}}_i = (\hat{\alpha}_i, \hat{\beta}_i, \hat{\tau}_i)^\mathsf{T} = g(\hat{\boldsymbol{\psi}}_i) = (\hat{\alpha}_i, \hat{\beta}_i, \log(\hat{\sigma}_i^2))^\mathsf{T}$$

as the data for lattice point i. Let H_{η_i} denote the Hessian matrix corresponding to the logarithm of the likelihood evaluated at the mode $\hat{\boldsymbol{\eta}}_i$, that is,

$$H_{\eta_i} = \nabla^2 \log(L(\boldsymbol{\eta}_i | \boldsymbol{y}_i))|_{\eta_i = \hat{\eta}_i}.$$

The observed information matrix, $\mathcal{I}_{\eta y_i}$, is equal to the negative Hessian matrix, $\mathcal{I}_{\eta y_i} = -H_{\eta_i}$, and it is given by

$$\mathcal{I}_{\eta y_i} = \begin{bmatrix} \exp(-\hat{\tau}_i)(F_i^\mathsf{T} F_i)_{1,1} & \exp(-\hat{\tau}_i)(F_i^\mathsf{T} F_i)_{1,2} & 0 \\ \exp(-\hat{\tau}_i)(F_i^\mathsf{T} F_i)_{2,1} & \exp(-\hat{\tau}_i)(F_i^\mathsf{T} F_i)_{2,2} & 0 \\ 0 & 0 & T/2 \end{bmatrix}.$$

The likelihood function for $\boldsymbol{\eta}_i$ and its approximation are such that

$$L(\boldsymbol{\eta}_i) = \pi(\boldsymbol{y}_i | \boldsymbol{\eta}_i) \approx c_i \hat{L}(\boldsymbol{\eta}_i) = c_i \pi(\boldsymbol{\eta}_i | \hat{\boldsymbol{\eta}}_i) = c_i \mathrm{N}(\boldsymbol{\eta}_i | \hat{\boldsymbol{\eta}}_i, \mathcal{I}_{\eta y_i}^{-1})$$

where c_i is a constant independent of $\boldsymbol{\eta}_i$. Now let $\boldsymbol{\eta} = (\boldsymbol{\alpha}^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\tau}^\mathsf{T})^\mathsf{T}$ with $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ as before, then the likelihood of $\boldsymbol{\eta}$, $L(\boldsymbol{\eta})$, is approximated with $\hat{L}(\boldsymbol{\eta})$, that is

$$L(\boldsymbol{\eta}) = \prod_{i=1}^{N} L(\boldsymbol{\eta}_i) \approx c \hat{L}(\boldsymbol{\eta}) = \prod_{i=1}^{N} c_i \hat{L}(\boldsymbol{\eta}_i) = c \mathrm{N}(\boldsymbol{\eta} | \hat{\boldsymbol{\eta}}, Q_{\eta y}^{-1})$$

where $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\alpha}}^\mathsf{T}, \hat{\boldsymbol{\beta}}^\mathsf{T}, \hat{\boldsymbol{\tau}}^\mathsf{T})^\mathsf{T}$ and $c = \prod_{i=1}^{N} c_i$. Define $\hat{\boldsymbol{\eta}}^* := (\hat{\alpha}_1, \hat{\beta}_1, \hat{\tau}_1, ..., \hat{\alpha}_N, \hat{\beta}_N, \hat{\tau}_N)^\mathsf{T}$ as a rearrangement of $\hat{\boldsymbol{\eta}}$ then

$$(Q_{\eta y}^*)^{-1} = \mathrm{bdiag}((\mathcal{I}_{\eta y_1})^{-1}, ..., (\mathcal{I}_{\eta y_N})^{-1}),$$

is known, and it is possible to rearrange $(Q_{\eta y}^*)^{-1}$ accordingly to get $Q_{\eta y}^{-1}$.

Here $\hat{\boldsymbol{\eta}}$ is taken as the data and $Q_{\eta y}^{-1}$ as fixed. The approximate data density is then

$$\pi(\hat{\boldsymbol{\eta}} | \boldsymbol{\eta}, \boldsymbol{\nu}, \theta) = \mathrm{N}(\hat{\boldsymbol{\eta}} | \boldsymbol{\eta}, Q_{\eta y}^{-1})$$

where $\boldsymbol{\nu}$ are all unknown parameters that are assigned a Gaussian prior distribution and $\boldsymbol{\theta}$ are all unknown hyperparameters. The vectors $\boldsymbol{x}$, $\boldsymbol{\eta}$, $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ are such that

$$\boldsymbol{x} = (\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\nu}^\mathsf{T})^\mathsf{T}, \quad \boldsymbol{\eta} = (\boldsymbol{\alpha}^\mathsf{T}, \boldsymbol{\beta}^\mathsf{T}, \boldsymbol{\tau}^\mathsf{T})^\mathsf{T}, \quad \boldsymbol{\nu} = (\boldsymbol{u}_\alpha^\mathsf{T}, \boldsymbol{u}_\beta^\mathsf{T}, \boldsymbol{u}_\tau^\mathsf{T})^\mathsf{T}$$

and

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_6)^\mathsf{T} = (\sigma_{u,\alpha}, \sigma_{\epsilon,\alpha}, \sigma_{u,\beta}, \sigma_{\epsilon,\beta}, \sigma_{u,\tau}, \sigma_{\epsilon,\tau})^\mathsf{T}.$$

To sample from the posterior density of $(\boldsymbol{x}, \boldsymbol{\theta})$, we first draw a sample from

$$\pi(\boldsymbol{\theta}|\hat{\boldsymbol{\eta}}) \propto \pi(\boldsymbol{\theta})\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\theta}) = \pi(\boldsymbol{\theta})\frac{\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{x}, \boldsymbol{\theta})\pi(\boldsymbol{x}|\boldsymbol{\theta})}{\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}}, \boldsymbol{\theta})},$$

and then we sample from the conditional posterior density of $\boldsymbol{x}$ given $\hat{\boldsymbol{\eta}}$ and $\boldsymbol{\theta}$. This conditional density, $\pi(\boldsymbol{x}|\hat{\boldsymbol{\eta}}, \boldsymbol{\theta})$, is specified in Section 1. Since the posterior density for $\boldsymbol{\theta}$ is such that the subvector $(\sigma_{u,\tau}, \sigma_{\epsilon,\tau})^\mathsf{T}$ is independent of the subvector $(\sigma_{u,\alpha}, \sigma_{\epsilon,\alpha}, \sigma_{u,\beta}, \sigma_{\epsilon,\beta})^\mathsf{T}$ then it is feasible to use grid sampling based on a two dimensional grid for $(\sigma_{u,\tau}, \sigma_{\epsilon,\tau})^\mathsf{T}$, and a four dimensional grid for $(\sigma_{u,\alpha}, \sigma_{\epsilon,\alpha}, \sigma_{u,\beta}, \sigma_{\epsilon,\beta})^\mathsf{T}$. This approach yields independent samples from the marginal posterior density of $\boldsymbol{\theta}$. Furthermore, as the samples from the full conditional posterior density of $\boldsymbol{x}$ are drawn from a Gaussian density then the sampling scheme yields independent samples from the approximated posterior density. As a result, there is no autocorrelation between the samples, and therefore, the proposed sampling scheme will be efficient.

Alternatively, a Metropolis step can be used to sample from the marginal posterior of $\boldsymbol{\theta}$. This requires the parameters to be defined on the real-line and calls for a transformation of the form $\boldsymbol{\kappa} = (\log\theta_1, ..., \log\theta_6)^\mathsf{T}$. Samples from $\pi(\boldsymbol{\kappa}|\hat{\boldsymbol{\eta}})$ are generated with a Metropolis step where the proposal distribution is

$$q(\boldsymbol{\kappa}^*|\boldsymbol{\kappa}^{(l)}, Q_\kappa^{-1}) = \mathrm{N}(\boldsymbol{\kappa}^*|\boldsymbol{\kappa}^{(l)}, Q_\kappa^{-1})$$

where $\boldsymbol{\kappa}^{(l)}$ is the value of $\boldsymbol{\kappa}$ in the l-th iteration. If the sampling scheme proposed by Roberts et al. (1997) is used then Q_κ is set equal to the negative Hessian matrix of $\log(\pi(\boldsymbol{\kappa}|\hat{\boldsymbol{\eta}}))$ times $\dim(\boldsymbol{\kappa})/2.382^2$. The prior density of each κ_k is

$$\pi(\kappa_k) = \lambda_k \exp\{-\lambda_k \exp(\kappa_k) + \kappa_k\}$$

where $k \in \{1, ..., 6\}$. The prior distribution of $\boldsymbol{\kappa}$ can now be written as

$$\pi(\boldsymbol{\kappa}) = \prod_{k=1}^{6} \pi(\kappa_k)$$

and the marginal posterior density of $\boldsymbol{\kappa}$ becomes

$$\pi(\boldsymbol{\kappa}|\hat{\boldsymbol{\eta}}) \propto \pi(\boldsymbol{\kappa})\frac{\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{x}, \boldsymbol{\kappa})\pi(\boldsymbol{x}|\boldsymbol{\kappa})}{\pi(\boldsymbol{x}|\boldsymbol{\kappa}, \hat{\boldsymbol{\eta}})}.$$

We can also apply Max-and-Smooth to infer $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ by approximating the likelihood function of the parameters of the i-th lattice point with a Gaussian density that has the same mean and variance as the normalized likelihood function of $(\alpha_i, \beta_i, \tau_i)^\mathsf{T}$. In the case of a normalized likelihood function stemming from a general Gaussian regression model with p regression coefficients in $\boldsymbol{\beta}$ and error log-variance τ, it can be shown that the covariance between $\boldsymbol{\beta}$ and τ is zero, thus, under the Gaussian approximation $\boldsymbol{\beta}$ and τ are independent. The marginal likelihood function of τ is a log-inverse-gamma density with parameters $\alpha = (T-p)/2$ and $\gamma = s^2(T-p)/2$ where

$$s^2 = (T-p)^{-1}\sum_{t=1}^{T}(y_t - \boldsymbol{f}_t\hat{\boldsymbol{\beta}})^2$$

and $\boldsymbol{f}_t$ contains the covariates of the t-th observation and it is also the t-th row of the design matrix F. The mean and variance of τ stemming from the normalized likelihood function are

$$\mathrm{E}(\tau) = \log(s^2) + \log((T-p)/2) - \psi((T-p)/2), \quad \mathrm{var}(\tau) = \psi'((T-p)/2).$$

Furthermore, the marginal likelihood function of $\boldsymbol{\beta}$ is a multivariate t-density with $T-p$ degrees of freedom and parameters

$$\boldsymbol{\mu} = (F^\mathsf{T}F)^{-1}F^\mathsf{T}\boldsymbol{y}, \quad \Sigma = s^2(F^\mathsf{T}F)^{-1}.$$

The mean and variance of $\boldsymbol{\beta}$ are

$$\mathrm{E}(\boldsymbol{\beta}) = (F^\mathsf{T}F)^{-1}F^\mathsf{T}\boldsymbol{y}, \quad \mathrm{var}(\boldsymbol{\beta}) = \frac{(T-p)}{(T-p-2)}s^2(F^\mathsf{T}F)^{-1}.$$

Therefore, the Gaussian approximation for $\boldsymbol{\beta}$ based on the mean and variance of the normalized likelihood function will have the same mean as the Gaussian approximation for $\boldsymbol{\beta}$ based on the ML estimates and the observed information while its variance is larger by a factor $T(T-p-2)^{-1}$.

Results

In this subsection we simulate data from a latent Gaussian model representing a linear regression models on a lattice. The purpose of this simulation is to evaluate the Gaussian approximation to the likelihood function, and to check whether the marginal posterior intervals capture the true values of the inferred parameters. The dimension of the lattice is 61×61. Figure 1 reflects the likelihood function of α, β and τ for a single lattice point with respect to the Gaussian approximations of the likelihood function based on the ML estimates and the observed information matrix, and based on the mean and variance of the normalized likelihood function. The part of the likelihood function corresponding to α, β and τ at the particular lattice point is normalized by integrating over the three variables. The normalized likelihood function is treated as a probability density and based on it the marginal probability functions of α, β and τ can be derived. The marginal normalized likelihood functions of α and β are scaled t-densities with $T-2$ degrees of freedom. The marginal normalized likelihood function of τ is a log-inverse-gamma density. This density is also referred to as a log-scaled-inverse χ^2 density with $T-2$ degrees of freedom. Figure 1 shows the marginal normalized likelihood functions of α, β and τ along with the marginal densities of the trivariate Gaussian densities that are used to approximated the likelihood function in the case of $T = 10, 20, 50$. The Gaussian approximation based on the ML estimates is reasonably good in the cases of α and β when $T = 20$ and $T = 50$ as expected since the marginal normalized likelihood functions are t-densities with 18 and 48 degrees of freedom. The Gaussian approximation is not as good for α and β when $T = 10$ as that case boils down to approximating a t-density with 8 degrees of freedom with a Gaussian density. In the case of the log-variance, τ, the Gaussian approximation is acceptable but it misses the right skewness of the marginal normalized likelihood function of τ. Figure 1 shows that

this skewness decreases rapidly as the sample size increases. A comparison of the results shown in Figure 1 reveals that the Gaussian approximation that has the same mean and variance as the normalized likelihood function gives a slightly better approximation than the Gaussian approximation that is based on the ML estimate and the observed information matrix. This is most obvious for τ when $T = 10$ and $T = 20$, but it is also visible for α and β when $T = 10$ and $T = 20$. However, the difference between the two Gaussian approximations to the normalized likelihood function is small when $T = 50$.

Figure 2 shows the marginal posterior densities of the hyperparameters and it can be seen that the marginal posterior densities capture the true values of $\sigma_{u,\alpha}$, $\sigma_{u,\beta}$ and $\sigma_{u,\tau}$ quite well. The marginal posterior densities of $\sigma_{\epsilon,\alpha}$, $\sigma_{\epsilon,\beta}$ and $\sigma_{\epsilon,\tau}$ in Figure 2 show that it is more difficult to determine the values of $\sigma_{\epsilon,\alpha}$, $\sigma_{\epsilon,\beta}$ and $\sigma_{\epsilon,\tau}$ compared to $\sigma_{u,\alpha}$, $\sigma_{u,\beta}$ and $\sigma_{u,\tau}$. In particular, the marginal posterior density of $\sigma_{\epsilon,\tau}$ gives negligible probability mass over the true value and the posterior mean is much larger than the true value.

Figure 3 shows 95% posterior credible intervals of the elements in $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ as a function of their true values as well as histograms of the difference between the posterior means and the true values when $T = 23$. Since the data are simulated, the coverage of the 95% posterior credible intervals is evaluated for each of the vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ by calculating the proportion of the credible intervals within each vector that capture the corresponding true values. The proportions for $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$ are 92.4%, 93.9% and 89.0%, respectively. The difference between the posterior means and the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are centred around zero while this same difference of the elements in $\boldsymbol{\tau}$ is centered around a value close to -0.1337. Thus, the posterior means are more likely to give an underestimate of the true value of the elements in $\boldsymbol{\tau}$. This could be due to the fact that the Gaussian approximation for the marginal normalized likelihood function based on the ML estimate for a given τ has a mean that is smaller than the mean of the actual marginal normalized likelihood function for τ. The Gaussian approximation based on the mean and variance of the normalized likelihood function gives mean that is greater than the mean of the Gaussian approximation based on the ML estimates by 0.1393 for all lattice points i. When the Gaussian approximation based on the mean and variance of the normalized likelihood function is applied (results not shown) then the bias defined above is close to zero, which supports using this approximation when the normalized likelihood function is skewed. This skewness decreases as the number of replicates within a group increases.

2.3 Linear regression model with temporally varying coefficients

Linear regression models for time referenced data can sometimes be improved by allowing the regression coefficients of the predictors to vary over time. Examples of data of this type are data on real estate prices (Gelfand et al., 2007) and spatio-temporal fillet yield data (Margeirsson et al., 2010). A general setup for models with temporally varying regression coefficients can be found in West and Harrison (1999) and Prado and West (2010).

Statistical model

Here $y_{t,k}$ denotes the k-th observation of the variable of interest at time t, and $\boldsymbol{x}_{t,k}$ is a vector of p predictors for the k-th observation at time t. The data within a given time period are assumed to follow the t-distribution with location parameter $\mu_{t,k} = \boldsymbol{x}_{t,k}^{\intercal}\boldsymbol{\beta}_t$, scale parameter σ_t and degrees of freedom parameter ϑ_t, i.e.,

$$y_{t,k} \sim t_{\vartheta_t}(\boldsymbol{x}_{t,k}^{\intercal}\boldsymbol{\beta}_t, \sigma_t^2), \quad \boldsymbol{\beta}_t \in \mathbb{R}^p, \quad \sigma_t > 0, \quad \vartheta_t > 0,$$

where $k \in \{1, ..., n_t\}$, $t \in \{1, ..., T\}$ and n_t is the number of observations at time t. Each time point t forms a group, so, here the number of groups is $G = T$. The data density for the data at time t, $\boldsymbol{y}_t = (y_{t,1}, ..., y_{t,n_t})^{\intercal}$, is given by

$$\pi(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \sigma_t, \vartheta_t) = \prod_{k=1}^{n_t} \frac{\Gamma((\vartheta_t+1)/2)}{\Gamma(\vartheta_t/2)\sqrt{\vartheta_t\pi}\sigma_t} \left\{1 + \frac{1}{\vartheta_t}\left(\frac{y_{t,k} - \boldsymbol{x}_{t,k}^{\intercal}\boldsymbol{\beta}_t}{\sigma_t}\right)^2\right\}^{-(\vartheta_t+1)/2}. \quad (2.11)$$

The parameters σ_t and ϑ_t are transformed to $\tau_t = \log(\sigma_t)$ and $\varphi_t = \log(\vartheta_t)$, so $\tau_t, \varphi_t \in \mathbb{R}$.

The parameters are collected into the vectors $\boldsymbol{\beta}_s = (\beta_{1,s}, ..., \beta_{T,s})^{\intercal}$, $s = 1, ..., p$, $\boldsymbol{\tau} = (\tau_1, ..., \tau_T)^{\intercal}$ and $\boldsymbol{\varphi} = (\varphi_1, ..., \varphi_T)^{\intercal}$. The latent level consists of linear predictors for these $p + 2$, vectors,

$$\begin{aligned} \boldsymbol{\beta}_s &= \boldsymbol{u}_{\beta,s} + \boldsymbol{\epsilon}_{\beta,s}, \quad s = 1, ..., p, \\ \boldsymbol{\tau} &= \boldsymbol{u}_{\tau} + \boldsymbol{\epsilon}_{\tau}, \\ \boldsymbol{\varphi} &= \boldsymbol{u}_{\varphi} + \boldsymbol{\epsilon}_{\varphi}, \end{aligned} \quad (2.12)$$

where $\boldsymbol{u}_{\beta,s}$, $\boldsymbol{u}_{\tau}$ and $\boldsymbol{u}_{\varphi}$ are modeled with random walk processes and $\boldsymbol{\epsilon}_{\beta,s}$, $\boldsymbol{\epsilon}_{\tau}$ and $\boldsymbol{\epsilon}_{\varphi}$ are unstructured random effects with standard deviations $\sigma_{\epsilon,\beta,s}$, $\sigma_{\epsilon,\tau}$ and $\sigma_{\epsilon,\varphi}$, respectively. For example, the model for $\boldsymbol{u}_{\tau}$ is such that $u_{t,\tau} = u_{t-1,\tau} + v_t$ where v_t are independent Gaussian random variables with mean zero and variance $\sigma_{u,\tau}^2$, and the precision matrix of $\boldsymbol{u}_{\tau}$ is $\sigma_{u,\tau}^{-2} R_T$ where R_T is specified by equation (2.9) above.

Inference

To infer the unknown parameters in the model, we apply Max-and-Smooth. Since the degrees of freedom parameter is difficult to infer, we propose using a prior density that is multiplied to the likelihood function. In particular, we propose using a log-gamma prior density with parameters α and γ. Its density is given by

$$\pi(\varphi_t) = \frac{\gamma^{\alpha}}{\Gamma(\alpha)} \exp\{\alpha\varphi_t - \gamma\exp(\varphi_t)\},$$

and its mean and variance are

$$\mathrm{E}(\varphi_t) = \psi(\alpha) - \log(\gamma), \quad \mathrm{var}(\varphi_t) = \psi'(\alpha).$$

Thus, the generalized likelihood function of $(\boldsymbol{\beta}_t, \tau_t, \varphi_t)$ is given by

$$L(\boldsymbol{\beta}_t, \tau_t, \varphi_t|\boldsymbol{y}_t) = \pi(\boldsymbol{y}_t|\boldsymbol{\beta}_t, \tau_t, \varphi_t)\pi(\varphi_t), \quad (2.13)$$

where σ_t and ϑ_t are replaced by $\exp(\tau_t)$ and $\exp(\varphi_t)$, respectively, in the data density for $\boldsymbol{y}_t$ given by (2.11). Here the maximum likelihood estimate of $(\beta_{t,1}, ..., \beta_{t,p}, \tau_t, \varphi_t)^\mathsf{T}$, evaluated by maximizing the above generalized likelihood function, are modeled as an LGM with $p+2$ linear predictors with the Gaussian–Gaussian model described in Section 3.3 of the main paper and the scheme in Section 1 is used for inference.

Let $\boldsymbol{\psi}_t = (\beta_{t,1}, ..., \beta_{t,p}, \sigma_t, \vartheta_t)^\mathsf{T}$ and let $g : \mathbb{R}^p \times \mathbb{R}^2_+ \to \mathbb{R}^{p+2}$ be a $(p+2)$-variate link function such that $g(x_1, ..., x_{p+2}) = (x_1, ..., x_p, \log(x_{p+1}), \log(x_{p+2}))^\mathsf{T}$. Then the inverse of g is $g^{-1}(u_1, ..., u_{p+2}) = (u_1, ..., u_p, \exp(u_{p+1}), \exp(u_{p+2}))^\mathsf{T}$. Furthermore, let

$$\boldsymbol{\eta}_t := (\beta_{t,1}, ..., \beta_{t,p}, \tau_t, \varphi_t)^\mathsf{T} = g(\boldsymbol{\psi_t}) = (\beta_{t,1}, ..., \beta_{t,p}, \log(\sigma_t), \log(\vartheta_t))^\mathsf{T}.$$

Now define $\hat{\boldsymbol{\eta}}_t := (\hat{\beta}_{t,1}, ..., \hat{\beta}_{t,p}, \hat{\tau}_t, \hat{\varphi}_t)^\mathsf{T}$ as the maximum likelihood estimate of the generalized likelihood function at time t given by (2.13) which is computed separately from the other time points. Instead of using $\boldsymbol{y}_t$ as the data, we treat $\hat{\boldsymbol{\eta}}_t$ as the data at time t. The Hessian matrix, corresponding to the logarithm of the generalized likelihood function evaluated at the mode $\hat{\boldsymbol{\eta}}_t$, is given by

$$H_{\eta_t} = \nabla^2 \log(L(\boldsymbol{\eta}_t|\boldsymbol{y}_t))|_{\eta_t = \hat{\eta}_t}.$$

The observed information matrix, $\mathcal{I}_{\eta y_t}$, is equal to the negative Hessian matrix, i.e., $\mathcal{I}_{\eta y_t} = -H_{\eta_t}$. The likelihood function for $\boldsymbol{\eta}_t$ and its approximation are such that

$$L(\boldsymbol{\eta}_t) = \pi(\boldsymbol{y}_t|\boldsymbol{\eta}_t) \approx c_t \hat{L}(\boldsymbol{\eta}_t) = c_t \pi(\boldsymbol{\eta}_t|\hat{\boldsymbol{\eta}}_t) = c_t \mathrm{N}(\boldsymbol{\eta}_t|\hat{\boldsymbol{\eta}}_t, \mathcal{I}^{-1}_{\eta y_t})$$

where c_t is a constant independent of $\boldsymbol{\eta}_t$. Now let $\boldsymbol{\eta} = (\boldsymbol{\beta}_1^\mathsf{T}, ..., \boldsymbol{\beta}_p^\mathsf{T}, \boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\varphi}^\mathsf{T})^\mathsf{T}$ with $\boldsymbol{\beta}_1$, ..., $\boldsymbol{\beta}_p$, $\boldsymbol{\tau}$ and $\boldsymbol{\varphi}$ as above, then the likelihood of $\boldsymbol{\eta}$, $L(\boldsymbol{\eta})$, is approximated with $\hat{L}(\boldsymbol{\eta})$, that is

$$L(\boldsymbol{\eta}) = \prod_{t=1}^{T} L(\boldsymbol{\eta}_t) \approx c\hat{L}(\boldsymbol{\eta}) = \prod_{t=1}^{T} c_t \hat{L}(\boldsymbol{\eta}_t) = c\mathrm{N}(\boldsymbol{\eta}|\hat{\boldsymbol{\eta}}, Q^{-1}_{\eta y})$$

where $\hat{\boldsymbol{\eta}} = (\hat{\boldsymbol{\beta}}_1^\mathsf{T}, ..., \hat{\boldsymbol{\beta}}_p^\mathsf{T}, \hat{\boldsymbol{\tau}}^\mathsf{T}, \hat{\boldsymbol{\varphi}}^\mathsf{T})^\mathsf{T}$ and $c = \prod_{t=1}^{T} c_t$. Define

$$\hat{\boldsymbol{\eta}}^* := (\hat{\beta}_{1,1}, ..., \hat{\beta}_{1,p}, \hat{\tau}_1, \hat{\varphi}_1, ..., \hat{\beta}_{T,1}, ..., \hat{\beta}_{T,p}, \hat{\tau}_T, \hat{\varphi}_T)^\mathsf{T}$$

as a rearrangement of $\hat{\boldsymbol{\eta}}$ then

$$(Q^*_{\eta y})^{-1} = \mathrm{bdiag}(\mathcal{I}^{-1}_{\eta y_1}, ..., \mathcal{I}^{-1}_{\eta y_T}),$$

is known, and it is possible to rearrange $(Q^*_{\eta y})^{-1}$ accordingly to get $Q^{-1}_{\eta y}$.

In the psuedo Gaussian–Gaussian model that is used for inference, $\hat{\boldsymbol{\eta}}$ is taken as the data and $Q^{-1}_{\eta y}$ as fixed. The approximate data distribution of $\hat{\boldsymbol{\eta}}$ is

$$\pi(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, \boldsymbol{\nu}, \boldsymbol{\theta}) = \mathrm{N}(\hat{\boldsymbol{\eta}}|\boldsymbol{\eta}, Q^{-1}_{\eta y})$$

where $\boldsymbol{\nu}$ are all unknown parameters that are assigned a Gaussian prior distribution and $\boldsymbol{\theta}$ are all unknown hyperparameters. The vectors $\boldsymbol{x}$, $\boldsymbol{\eta}$, $\boldsymbol{\nu}$ and $\boldsymbol{\theta}$ are such that

$$\boldsymbol{x} = (\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\nu}^\mathsf{T})^\mathsf{T}, \quad \boldsymbol{\eta} = (\boldsymbol{\beta}_1^\mathsf{T}, ..., \boldsymbol{\beta}_p^\mathsf{T}, \boldsymbol{\tau}^\mathsf{T}, \boldsymbol{\varphi}^\mathsf{T})^\mathsf{T}, \quad \boldsymbol{\nu} = (\boldsymbol{u}^\mathsf{T}_{\beta,1}, ..., \boldsymbol{u}^\mathsf{T}_{\beta,p}, \boldsymbol{u}^\mathsf{T}_{\tau}, \boldsymbol{u}^\mathsf{T}_{\varphi})^\mathsf{T}$$

and

$$\boldsymbol{\theta} = (\theta_1, ..., \theta_{2p+4})^{\mathsf{T}} = (\sigma_{u,\beta,1}, \sigma_{\epsilon,\beta,1}, ..., \sigma_{u,\beta,p}, \sigma_{\epsilon,\beta,p}, \sigma_{u,\tau}, \sigma_{\epsilon,\tau}, \sigma_{u,\varphi}, \sigma_{\epsilon,\varphi})^{\mathsf{T}}.$$

Posterior samples are generated with the sampling schemes described in Section 1, however, further details will be omitted.

Results

A simulation study was conducted using the model above with $p = 1$, so, at time t the parameters are $\boldsymbol{\eta}_t = (\beta_{t,1}, \tau_t, \varphi_t)^{\mathsf{T}}$. The likelihood function of $(\beta_{t,1}, \tau_t, \varphi_t)^{\mathsf{T}}$ for a particular time point was explored. The true values of the parameters in this simulation were $\beta_{t,1} = 10$, $\tau_t = 0$ ($\sigma_t = 1$) and $\varphi_t = 2.0794$ ($\vartheta_t = 8$). The degrees of freedom parameter can be difficult to infer. To regulate the log-degrees of freedom parameter, φ_t, we opt for a log-gamma prior density. Assume that according to prior knowledge the degrees of freedom are in the interval from 1 to 30. Thus, we select a log-gamma prior density with parameters $\alpha = 2$ and $\gamma = 0.2$, which has 95% of its probability mass over the interval $(0.1915, 3.327)$. In terms of the degrees of freedom parameter, ϑ_t, this translates to the interval $(1.211, 27.858)$.

Figure 4 shows the normalized generalized likelihood functions through the marginal normalized generalized likelihood functions of $\beta_{t,1}$, $\tau_t = \log(\sigma_t)$ and $\varphi_t = \log(\vartheta_t)$ when the sample size at time t, n_t, is 20, 40 and 80. The Gaussian approximation based on the ML estimates and the observed information matrix is also shown in Figure 4 in terms of the marginal Gaussian approximations of $\beta_{t,1}$, τ_t and φ_t. The Gaussian approximation for $\beta_{t,1}$ is very good for the three sample sizes and for τ_t and φ_t when $n_t = 40, 80$. The Gaussian approximation is reasonably good for τ_t and φ_t when $n_t = 20$, however, a small degree of skewness is apparent in the two marginal generalized likelihood functions in this case. This skewness vanishes as the sample size increases. The generalized likelihood function is influenced by the selected log-gamma prior density for φ_t. The prior density for φ_t can be useful for regularizing the inference for the unknown parameters when the sample size is 80 or less since in these cases the unregularized likelihood function provides a limited amount of information about the degrees of freedom parameter which affects the information about the other parameters.

2.4 Spatio-temporal model for count data

Statistical model

In this section $y_{i,t}$ denotes the observed count at lattice point i and time t. We assume that the counts follow a Poisson distribution, and that $y_{i,t}$ has mean $\lambda_{i,t}$, i.e.,

$$y_{i,t} \sim \text{Poisson}(\lambda_{i,t}), \quad \lambda_{i,t} > 0,$$

where $t \in \{1, ..., T\}$, T is the number of time points, $i \in \{1, ..., N\}$, N is the number of lattice points, and the lattice has dimensions n_1 and n_2, so, $N = n_1 n_2$. The spatio-temporal cell at lattice point i and time t forms a group, so, the total number of

spatio-temporal groups is $G = NT$. The mean of $y_{i,t}$ is transformed to $\eta_{i,t} = \log(\lambda_{i,t})$, $\eta_{i,t} \in \mathbb{R}$. The data density of $y_{i,t}$ is given by

$$\pi(y_{i,t}|\eta_{i,t}) = \frac{1}{y_{i,t}!} \exp\{\eta_{i,t} y_{i,t} - \exp(\eta_{i,t})\}.$$

The parameters are collected into vectors by grouping first over the lattice points, i.e., $\boldsymbol{\eta}_t = (\eta_{1,t}, ..., \eta_{N,t})^{\mathsf{T}}$, $t \in \{1, ..., T\}$. The latent level consists of the following model

$$\boldsymbol{\eta}_t = \boldsymbol{u}_t + \boldsymbol{\epsilon}_t, \quad \boldsymbol{u}_{t+1} = \boldsymbol{u}_t + \boldsymbol{v}_t,$$

so, $\boldsymbol{u}_t$ is modeled as a vector random walk where the vector $\boldsymbol{v}_t$ is independent of $\boldsymbol{u}_t$, and $\boldsymbol{v}_t$ modeled as a Gaussian random vector with mean zero and precision matrix $\sigma_{u,\eta}^{-2} Q_u$ where Q_u is given by equation (2.8) and $\sigma_{u,\eta}^2$ is an unknown variance parameter. In fact, $\boldsymbol{v}_t$ is modeled as a first-order IGMRF (Rue and Held, 2005, Section 3.3.2, pp. 104–108), and $\sigma_{u,\eta}^2/4$ is the conditional variance of an element in $\boldsymbol{v}_t$ located at an interior point of the lattice.

Inference

Max-and-Smooth is applied to infer $\boldsymbol{\eta}$ and $\sigma_{u,\eta}$. To stabilize the likelihood function of $\eta_{i,t}$ in the case of a small $y_{i,t}$, it is multiplied by a log-gamma prior density with parameters α and γ, namely,

$$\pi(\eta_{i,t}) = \frac{\gamma^\alpha}{\Gamma(\alpha)} \exp\{\alpha \eta_{i,t} - \gamma \exp(\eta_{i,t})\}.$$

The mean and variance of this density are $\psi(\alpha) - \log(\gamma)$ and $\psi'(\alpha)$, respectively. The corresponding normalized generalized likelihood function is

$$L(\eta_{i,t}) = \frac{(\gamma+1)^{\alpha+y_{i,t}}}{\Gamma(\alpha + y_{i,t})} \exp\{(\alpha + y_{i,t})\eta_{i,t} - (\gamma+1)\exp(\eta_{i,t})\},$$

and it has mean and variance

$$\tilde{\eta}_{i,t} = \psi(\alpha + y_{i,t}) - \log(\gamma + 1), \quad \omega_{\eta,i,t}^2 = \psi'(\alpha + y_{i,t}).$$

The mode and inverse observed information of the normalized generalized likelihood function are

$$\hat{\eta}_{i,t} = \log(\alpha + y_{i,t}) - \log(\gamma + 1), \quad \sigma_{\eta,i,t}^2 = (\alpha + y_{i,t})^{-1}.$$

Thus, the two Gaussian approximations of the generalized likelihood function are,

$$\hat{L}(\eta_{i,t}) = \mathrm{N}(\eta_{i,t}|\hat{\eta}_{i,t}, \sigma_{\eta,i,t}^2), \quad \tilde{L}(\eta_{i,t}) = \mathrm{N}(\eta_{i,t}|\tilde{\eta}_{i,t}, \omega_{\eta,i,t}^2).$$

In the case of $y_{i,t} > 0$ the normalized likelihood function and its two Gaussian approximations are a special case of the above formulas, namely, the values of α and γ are both set equal to zero. The normalized likelihood function is

$$L(\eta_{i,t}) = \frac{1}{\Gamma(y_{i,t})} \exp\{y_{i,t}\eta_{i,t} - \exp(\eta_{i,t})\},$$

and it has mean and variance

$$\tilde{\eta}_{i,t} = \psi(y_{i,t}), \quad \omega^2_{\eta,i,t} = \psi'(y_{i,t}).$$

The mode and inverse observed information of the normalized likelihood function are

$$\hat{\eta}_{i,t} = \log(y_{i,t}), \quad \sigma^2_{\eta,i,t} = (y_{i,t})^{-1}.$$

The two Gaussian approximations of the normalized likelihood function are specified with the means and variances above.

Now let $\boldsymbol{\eta} = (\boldsymbol{\eta}_1^\mathsf{T}, ..., \boldsymbol{\eta}_T^\mathsf{T})^\mathsf{T}$, and let $L(\boldsymbol{\eta})$ denote either the likelihood function or the generalized likelihood function of $\boldsymbol{\eta}$. $L(\boldsymbol{\eta})$ is approximated with either $\hat{L}(\boldsymbol{\eta})$ or $\tilde{L}(\boldsymbol{\eta})$ where

$$L(\boldsymbol{\eta}) = \prod_{t=1}^{T}\prod_{i=1}^{N} L(\eta_{i,t}) \approx c\hat{L}(\boldsymbol{\eta}) = \prod_{t=1}^{T}\prod_{i=1}^{N} c_{i,t}\hat{L}(\eta_{i,t}) = \prod_{t=1}^{T}\prod_{i=1}^{N} c_{i,t}\mathrm{N}(\eta_{i,t}|\hat{\eta}_{i,t}, \sigma^2_{\eta,i,t})$$

and

$$L(\boldsymbol{\eta}) = \prod_{t=1}^{T}\prod_{i=1}^{N} L(\eta_{i,t}) \approx c\tilde{L}(\boldsymbol{\eta}) = \prod_{t=1}^{T}\prod_{i=1}^{N} c_{i,t}\tilde{L}(\eta_{i,t}) = \prod_{t=1}^{T}\prod_{i=1}^{N} c_{i,t}\mathrm{N}(\eta_{i,t}|\tilde{\eta}_{i,t}, \omega^2_{\eta,i,t}).$$

The Gaussian–Gaussian pseudo model treats either $\hat{\eta}_{i,t}$ or $\tilde{\eta}_{i,t}$ as the data, and the corresponding known error variances are $\sigma^2_{\eta,i,t}$ or $\omega^2_{\eta,i,t}$ respectively. The approximate data densities of $\hat{\eta}_{i,t}$ and $\tilde{\eta}_{i,t}$ are

$$\pi(\hat{\eta}_{i,t}|\eta_{i,t}, \boldsymbol{\nu}, \theta) = \mathrm{N}(\hat{\eta}_{i,t}|\eta_{i,t}, \sigma^2_{\eta,i,t}), \quad \pi(\tilde{\eta}_{i,t}|\eta_{i,t}, \boldsymbol{\nu}, \theta) = \mathrm{N}(\tilde{\eta}_{i,t}|\eta_{i,t}, \omega^2_{\eta,i,t}),$$

for $i = 1, ..., N$, $t = 1, ..., T$, where $\boldsymbol{\nu}$ are all the unknown parameters that are assigned a Gaussian prior distribution and θ is the unknown hyperparameter. The vectors $\boldsymbol{x}$ and $\boldsymbol{\nu}$ are such that

$$\boldsymbol{x} = (\boldsymbol{\eta}^\mathsf{T}, \boldsymbol{\nu}^\mathsf{T})^\mathsf{T}, \quad \boldsymbol{\nu} = (\boldsymbol{u}_1^\mathsf{T}, ..., \boldsymbol{u}_T^\mathsf{T})^\mathsf{T},$$

where $\boldsymbol{\eta}$ is as above, and $\theta = \sigma_{u,\eta}$. Max-and-Smooth is used to infer the unknown parameters, see Section 1 and in Section 3.3 in the main paper. Further details will not be specified here.

Results

Here we look at two scenarios. The first scenario is such that we assume that the number of counts is small, i.e., $y_{i,t} = 0$, $y_{i,t} = 1$ and $y_{i,t} = 2$. For the first scenario we select a log-gamma prior density for $\eta_{i,t}$ with parameters $\alpha = 2$ and $\gamma = 8$. The product of the prior density and the likelihood function forms the generalized likelihood function of $\eta_{i,t}$. This prior density for $\eta_{i,t}$ is such that 95% of its probability mass is over the interval $(-3.4974, -0.3618)$ which translates to the interval $(0.03028, 0.69646)$ in terms of $\lambda_{i,t}$. This is an informative prior and should only be used if there is a priori high

certainty of $\lambda_{i,t}$ taking values in this interval. Under the second scenario larger counts are expected, in particular, we assume that observed counts between 5 and 100 are expected. Furthermore, the normalized likelihood function is used, not the normalized generalized likelihood function.

Figure 5 shows the normalized generalized likelihood function of $\eta_{i,t}$ when the observed value is $y_{i,t} = 0, 1, 2$, and the corresponding Gaussian approximations based on the generalized ML estimates and the inverse of the observed information, and the one based on the mean and variance of the normalized generalized likelihood function. In the case of all the three values of $y_{i,t}$, the Gaussian approximation based the mean and variance of the normalized generalized likelihood function is giving a better match since it captures better the variability of the true normalized generalized likelihood function and is closer to its tails than the Gaussian approximation based on the generalized ML estimates.

Figure 6 shows results for the second scenario using the same setup as in Figure 5. Under this scenario the observed values of $y_{i,t}$ are equal to 10, 50 and 90, and the normalized likelihood function is explored as opposed to the normalized generalized likelihood function. The two Gaussian approximations match the normalized generalized likelihood function very well in the case of $y_{i,t} = 50$ and $y_{i,t} = 90$. In the case of $y_{i,t} = 10$ the match is reasonably good for both of the Gaussian approximations. However, it is slightly better in the case of the Gaussian approximation with the mean and variance of the normalized likelihood function which has mean and standard deviation 2.252 and 0.3243, respectively, while the Gaussian approximation based on the ML estimates has mean and standard deviation 2.303 and 0.3162, respectively.

3 Block bootstrapping of score differences in the meteorological application

This section supplements Section 5 of the main manuscript. To further assess statistical significance of the improvement of SPAT1 over MLE, we now describe a basic uncertainty assessment of the difference in CRPS between these two approaches. Specifically, we approximated the sampling distribution of the average CRPS difference between the MLE and SPAT1 predictions using block bootstrapping. We treat all data as independent in time because the time increment between successive samples is one year, long enough to ignore the temporal correlation in small scale meteorological settings. To account for spatial correlation, we divided the spatial domain into a number of S rectangular non-overlapping blocks. This yields a total of $23 \times S$ blocks. We resample these blocks $23 \times S$ times with replacement, and then create a bootstrap sample of the average CRPS difference. The bootstrap distribution was estimated from 500 replicates. When using $S = 3$ spatial blocks (i.e. assuming only 3 effective spatial degrees of freedom) we obtain a bootstrap mean of 0.0025 and bootstrap standard deviation of 0.0013 for the CRPS difference, and a bootstrap p-value of 0.032. For $S = 10$ spatial degrees of freedom, we obtain a bootstrap mean of 0.0025, bootstrap standard deviation of 0.0012, and a bootstrap p-value of 0.006. For $S > 10$ the bootstrap p-value is essentially zero. Based on visual inspection of maps of CRPS differences, a conservative estimate of the

spatial degrees of freedom is at least 20, so the improvement of SPAT1 over MLE, albeit small, can thus be considered statistically significant with high confidence.

While this simple approach based on bootstrap p-values is not really in line with the classical Bayesian philosophy, a similar approach was used in Bakka et al. (2018) and Bakka et al. (2019), and we think it still provides valuable information about the significant superiority of SPAT1 over MLE. For a more rigorous Bayesian predictive and uncertainty assessment, see, e.g., Rubin (1981) who introduces the Bayesian bootstrap, Leininger and Gelfand (2017) who discuss methods for Bayesian inference and model assessment (though here, in the context of point patterns), and Krüger et al. (2020) who discuss using the CRPS in the Bayesian context.

References

Bakka, H., Castro-Camilo, D., Franco-Villoria, M., Freni-Sterrantino, A., Huser, R., Opitz, T., and Rue, H. (2018). "Discussion of 'Using stacking to average Bayesian predictive distributions' by Yao et al." *Bayesian Analysis*, 13: 917–1003. 20

Bakka, H., Vanhatalo, J., Illian, J. B., Simpson, D. P., and Rue, H. (2019). "Non-stationary Gaussian models with physical barriers." *Spatial Statistics*, 29: 268–288. 20

Gelfand, A. E., Banerjee, S., Sirmans, C., Tu, Y., and Ong, S. E. (2007). "Multilevel modeling using spatial processes: Application to the Singapore housing market." *Computational Statistics & Data Analysis*, 51(7): 3567–3579. 13

Krüger, F., Lerch, S., Thorarinsdottir, T. L., and Gneiting, T. (2020). "Predictive inference based on Markov chain Monte Carlo output." arXiv preprint 1608.06802. 20

Leininger, T. J. and Gelfand, A. E. (2017). "Bayesian inference and model assessment for spatial point patterns using posterior predictive samples." *Bayesian Analysis*, 12: 1–30. 20

Margeirsson, S., Hrafnkelsson, B., Jónsson, G. R., Jensson, P., and Arason, S. (2010). "Decision making in the cod industry based on recording and analysis of value chain data." *Journal of Food Engineering*, 99(2): 151–158. 13

Prado, R. and West, M. (2010). *Time Series: Modeling, Computation, and Inference*. CRC Press. 13

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). "Weak convergence and optimal scaling of random walk Metropolis algorithms." *The Annals of Applied Probability*, 7(1): 110–120. 11

Rubin, D. B. (1981). "The Bayesian bootstrap." *Annals of Statistics*, 9: 130–134. 20

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. CRC press. 5, 8, 17

Simpson, D., Rue, H., Riebler, A., Martins, T. G., Sørbye, S. H., et al. (2017). "Penal-

ising model component complexity: A principled, practical approach to constructing priors." *Statistical Science*, 32(1): 1–28. 9

West, M. and Harrison, J. (1999). *Bayesian Forecasting and Dynamic Models, Second Edition*. Springer Science & Business Media. 13

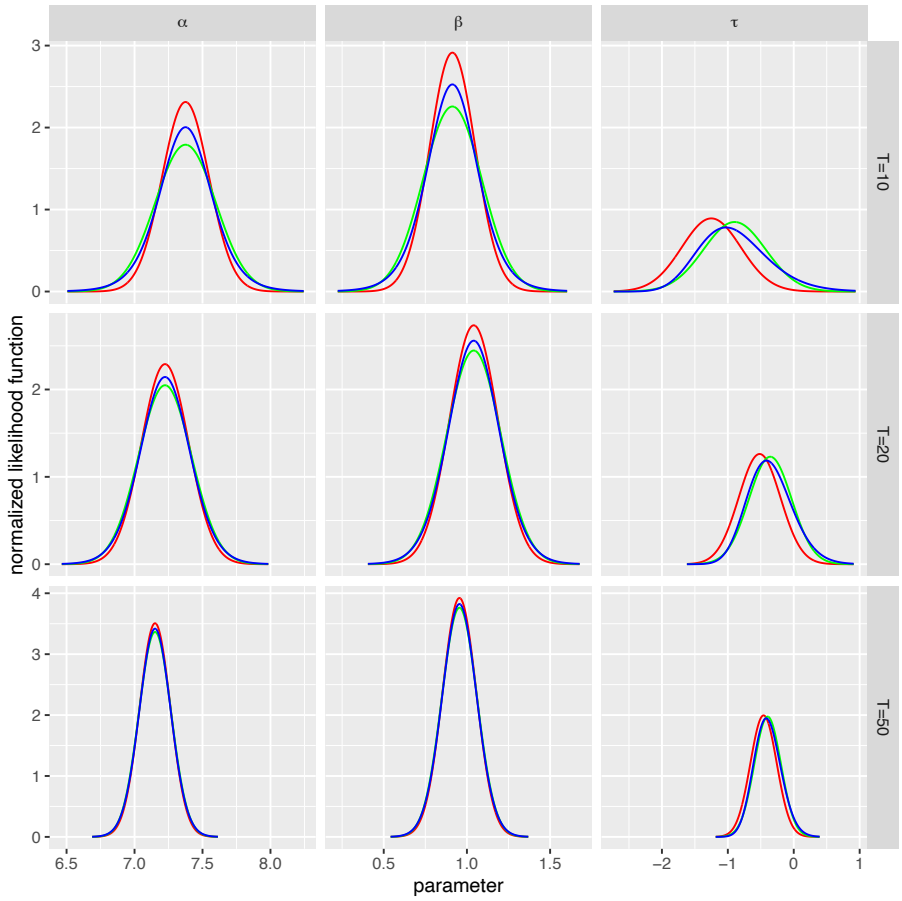

Figure 1: Linear regression on a lattice. The marginal normalized likelihood functions (blue lines) of α (left column of panel), β (middle column of panel) and τ (right column of panel), along with the marginal densities of the Gaussian approximations based on (a) the ML estimates and the inverse of the observed information (red lines), and (b) the mean and variance of the normalized likelihood function (green lines). The top, middle and bottom row panels show the results for $T = 10, 20, 50$, respectively.

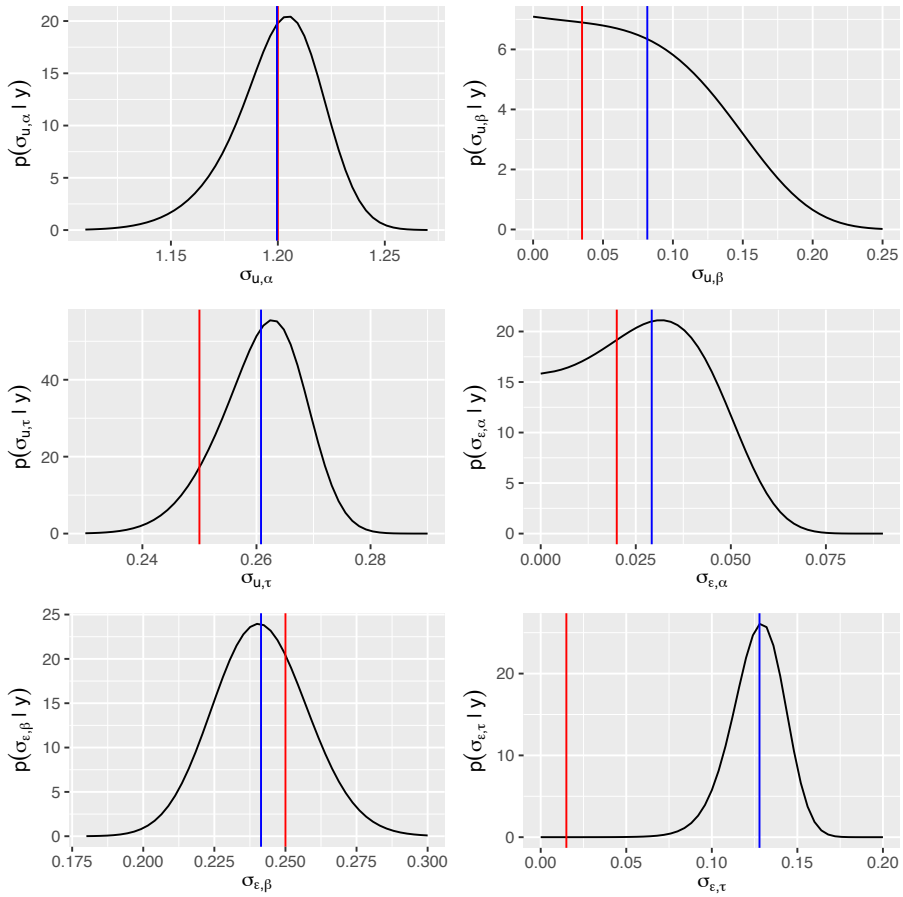

Figure 2: Linear regression on a lattice. The marginal posterior densities of the hyper-parameters (black lines). Top panel: $\sigma_{u,\alpha}$ (left), $\sigma_{\epsilon,\alpha}$ (right). Middle panel: $\sigma_{u,\beta}$ (left), $\sigma_{\epsilon,\beta}$ (right). Bottom panel: $\sigma_{u,\tau}$ (left), $\sigma_{\epsilon,\tau}$ (right). The solid vertical red lines show the true values while the solid vertical blue line show the posterior means. In the case of $\sigma_{u,\alpha}$, the true value and the posterior mean are close.

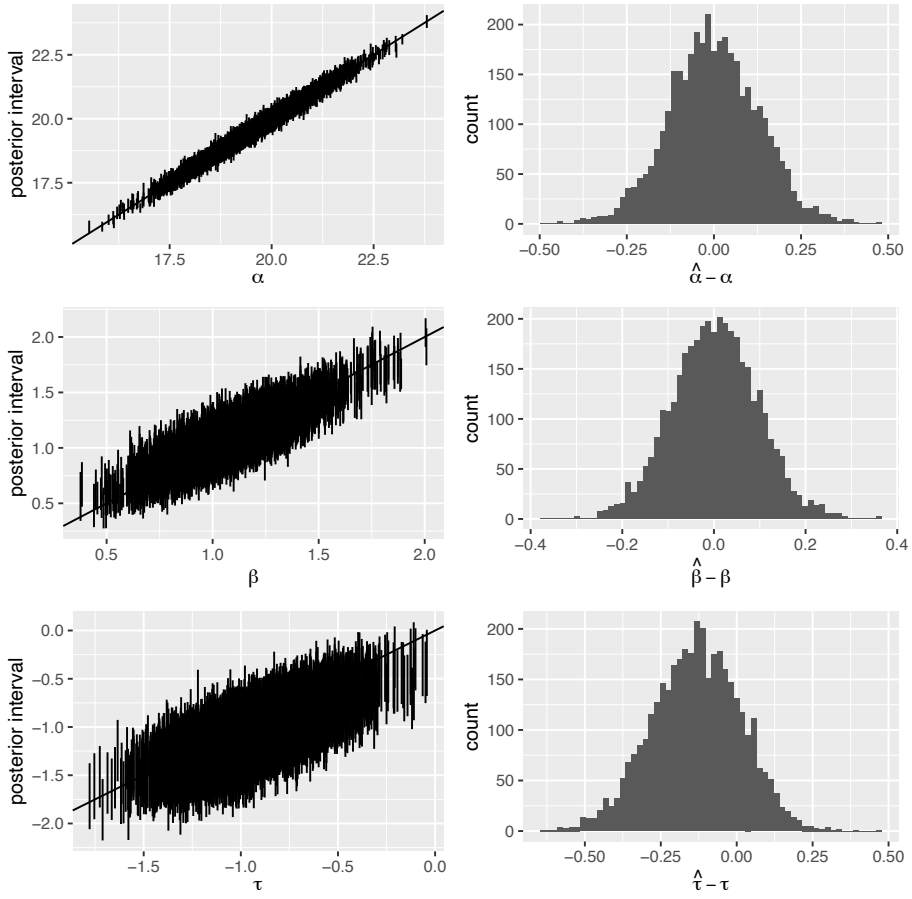

Figure 3: Linear regression on a lattice. The elements in $\boldsymbol{\alpha}$ (top row of panel): the 95% posterior intervals versus the true values (left column of panel) and a histogram of the posterior means minus the true values (right column of panel). The middle row and the bottom row of the panel show the same for the elements in $\boldsymbol{\beta}$ and the elements in $\boldsymbol{\tau}$, respectively.

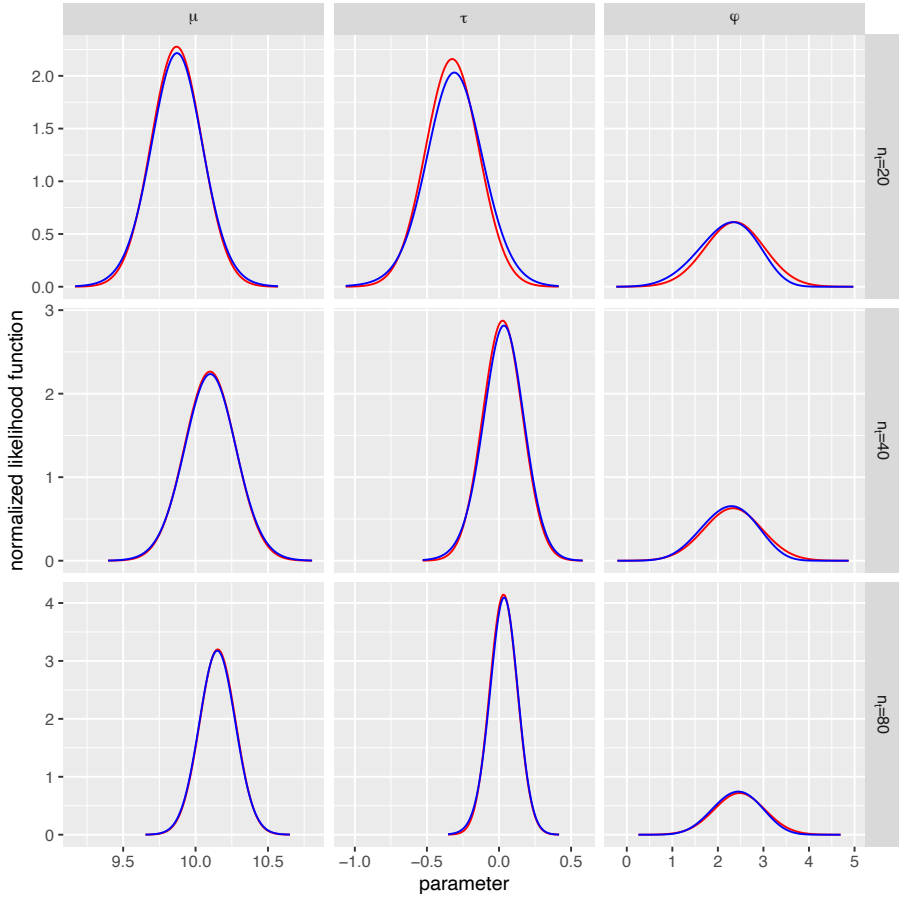

Figure 4: Linear regression model with temporally varying regression coefficients. In this model the error terms follow a t-distribution, and its scale and degree of freedom parameters also vary with time. The marginal normalized generalized likelihood functions (blue lines) of $\beta_{t,1}$ (left column of panel), $\tau_t = \log(\sigma_t)$ (middle column of panel) and $\varphi_t = \log(\vartheta_t)$ (right column of panel) are shown along with the marginal Gaussian approximations based on the ML estimates and the observed information matrix (red lines). The top row, middle row and bottom row of the panel show the results for $n_t = 20, 40, 80$, respectively.

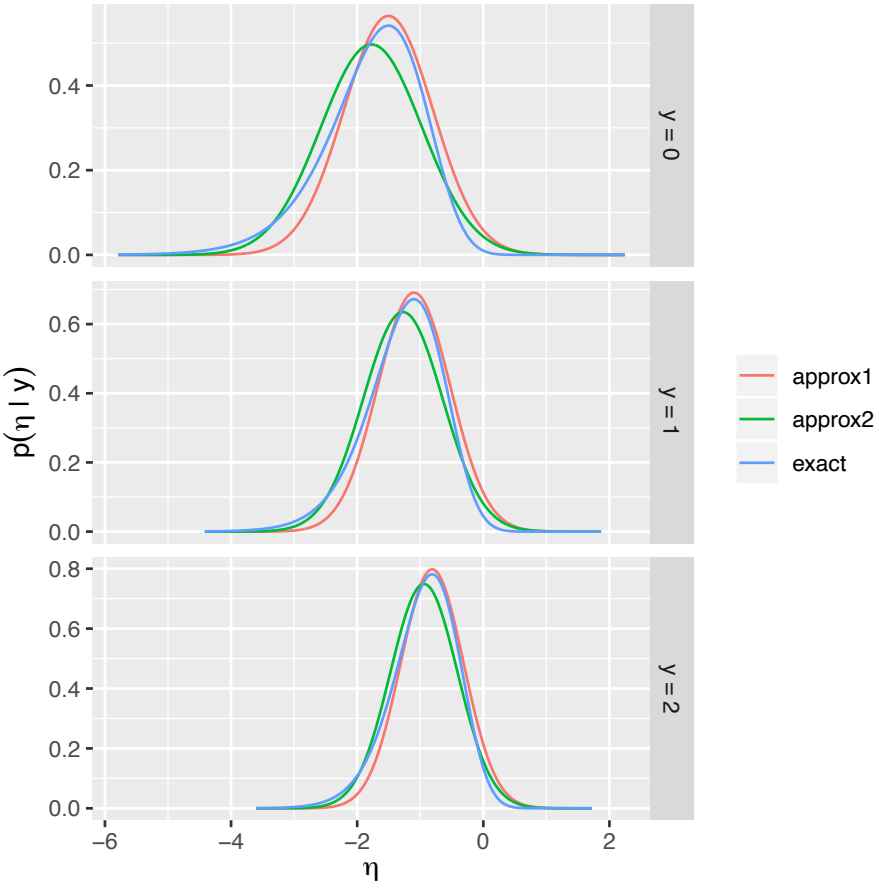

Figure 5: Spatio-temporal model for count data. The normalized generalized likelihood functions of η (blue lines) plotted along with the Gaussian approximations based on (a) the generalized ML estimates and the inverse of the observed information (red lines), and (b) the mean and variance of the normalized generalized likelihood function (green lines). The top row, the middle row and the bottom row of the panel show the results for $y = 0, 1, 2$, respectively. The likelihood function is multiplied by a log-gamma prior density with parameters $\alpha = 2$ and $\gamma = 8$.

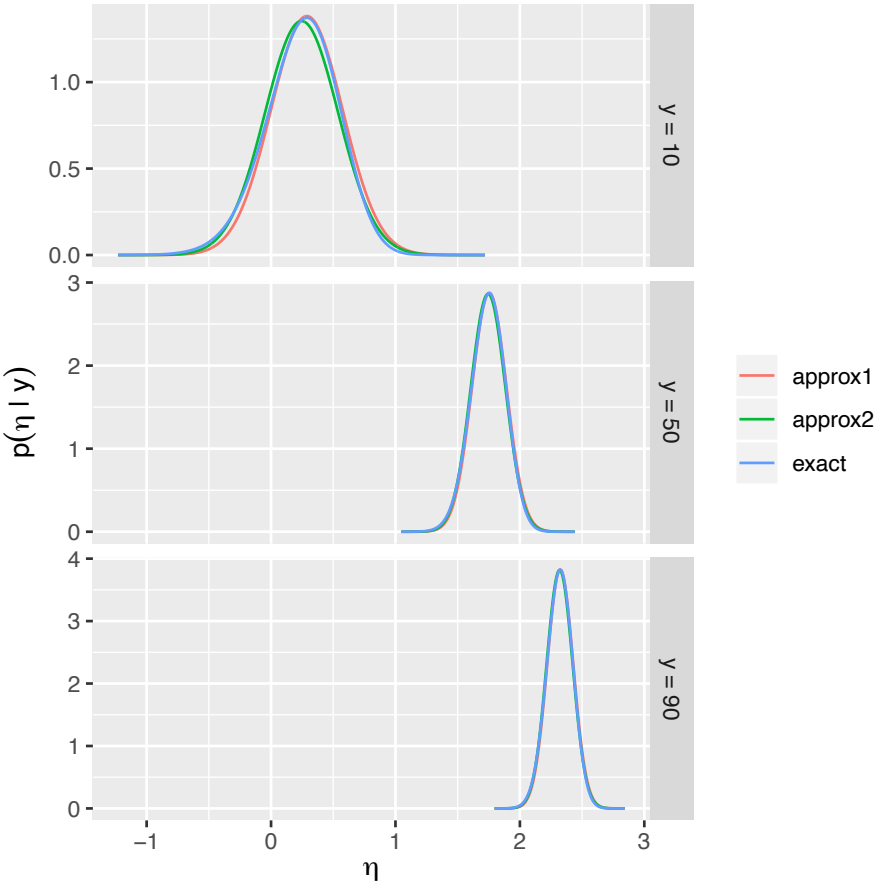

Figure 6: Spatio-temporal model for count data. The normalized likelihood functions of η (blue lines) plotted along with the Gaussian approximations based on (a) the ML estimates and the inverse of the observed information (red lines), and (b) the mean and variance of the normalized likelihood function (green lines). The top row, the middle row and the bottom row of the panel show the results for $y = 10, 50, 90$, respectively.