Fitting Conditional and Simultaneous Autoregressive Spatial Models in hglm

by Moudud Alam, Lars Rönnegård, and Xia Shen

Abstract We present a new version ($\geqslant$ 2.0) of the **hglm** package for fitting hierarchical generalized linear models (HGLMs) with spatially correlated random effects. `CAR()` and `SAR()` families for conditional and simultaneous autoregressive random effects were implemented. Eigen decomposition of the matrix describing the spatial structure (e.g., the neighborhood matrix) was used to transform the CAR/SAR random effects into an independent, but heteroscedastic, Gaussian random effect. A linear predictor is fitted for the random effect variance to estimate the parameters in the CAR and SAR models. This gives a computationally efficient algorithm for moderately sized problems.

Introduction

We present an algorithm for fitting spatial generalized linear models with conditional and simultaneous autoregressive (CAR & SAR; Besag, 1974; Cressie, 1993) random effects. The algorithm completely avoids the need to differentiate the spatial correlation matrix by transforming the model using an eigen decomposition of the precision matrix. This enables the use of already existing methods for hierarchical generalized linear models (HGLMs; Lee and Nelder, 1996) and this algorithm is implemented in the latest version ($\geqslant$ 2.0) of the **hglm** package (Rönnegård et al., 2010).

The **hglm** package, up to version 1.2-8, provides the functionality for fitting HGLMs with uncorrelated random effects using an extended quasi likelihood method (EQL; Lee et al., 2006). The new version includes a first order correction of the fixed effects based on the current EQL fitting algorithm, which is more precise than EQL for models having non-normal outcomes (Lee and Lee, 2012). Similar to the h-likelihood correction of Noh and Lee (2007), it corrects the estimates of the fixed effects and thereby also reduces potential bias in the estimates of the dispersion parameters for the random effects (variance components). The improvement in terms of reduced bias is substantial for models with a large number of levels in the random effect (Noh and Lee, 2007), which is often the case for spatial generalized linear mixed models (GLMMs).

Earlier versions of the package allow modeling of the dispersion parameter of the conditional mean model, with fixed effects, but not the dispersion parameter(s) of the random effects. The current implementation, however, enables the user to specify a linear predictor for each dispersion parameter of the random effects. Adding this option was a natural extension to the package because it allowed implementation of our proposed algorithm that fits CAR and SAR random effects.

Though the CAR/SAR models are widely used for spatial data analysis there are not many software packages which can be used for their model fitting. GLMMs with CAR/SAR random effects are often fitted in a Bayesian way using BUGS software (e.g., WinBUGS; Lunn et al., 2000, or alike), which leads to extremely slow computation due to its dependence on Markov chain Monte Carlo simulations. The R package **INLA** (Martins et al., 2013) provides a relatively fast Bayesian computation of spatial HGLMs via Laplace approximation of the posterior distribution where the model development has focused on continuous domain spatial modeling (Lindgren and Rue, 2015) whereas less focus has been on discrete models including CAR and SAR. Recently, package **spaMM** (Rousset and Ferdy, 2014) was developed to fit spatial HGLMs but is rather slow even for moderately sized data. Here, we extend package **hglm** to also provide fast computation of HGLMs with CAR and SAR random effects and at the same time an attempt has been made to improve the accuracy of the estimates by including corrections for the fixed effects.

The rest of the paper is organized as follows. First we give an introduction to CAR and SAR structures and present the h-likelihood theory together with the eigen decomposition of the covariance matrix of the CAR random effects and show how it simplifies the computation of the model. Then we present the R code implementation, show the use of the implementation for two real data examples and evaluate the package using simulations. The last section concludes the article.

CAR and SAR structures in HGLMs

HGLMs with spatially correlated random effects are commonly used in spatial data analysis (Cressie, 1993; Wall, 2004). A spatial HGLM with Gaussian CAR random effects is given by

$$\begin{array}{lcll} E\left(z_s|u_s\right) & = & \mu_s, & s=1,2,\ldots,n, \\ g\left(\mu_s\right) = \eta_s & = & \mathbf{X}_s\boldsymbol{\beta}+\mathbf{Z}_s u_s, & \end{array} \tag{1}$$

$$z_s|u_s \sim Exponential\ Family, \tag{2}$$

with $z_s|u_s \perp z_t|u_t, \forall s \neq t$ and

$$\mathbf{u} = \left(u_1, u_2, \ldots, u_n\right)^T \sim N\left(\mathbf{0}, \Sigma = \tau\left(\mathbf{I}-\rho\mathbf{D}\right)^{-1}\right), \tag{3}$$

where s represents a location identified by the coordinates $\left(x\left(s\right), y\left(s\right)\right)$, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{X}_s$ is the fixed-effect design matrix for location s, u_s being the location specific random effects and $\mathbf{Z}_s$ is a design matrix associated with u_s. While

$$\mathbf{u} = \left(u_1, u_2, \ldots, u_n\right)^T \sim N\left(\mathbf{0}, \Sigma = \tau\left(\mathbf{I}-\rho\mathbf{D}\right)^{-1}\left(\mathbf{I}-\rho\mathbf{D}^T\right)^{-1}\right) \tag{4}$$

gives a Gaussian SAR random effects structure. The $\mathbf{D}$ matrix in Equations (3) and (4) is some known function of the location coordinates (see, e.g., Clayton and Kaldor, 1987) and ρ is often referred to as the spatial dependence parameter (Hodges, 2013). In application, $\mathbf{D}$ is often a neighborhood matrix whose diagonal elements are all 0, and off-diagonal elements (s,t) are 1 if locations s and t are neighbors.

If $\boldsymbol{\Sigma} = \tau\mathbf{I}$, i.e., there is no spatial correlation, then this model may be estimated by using usual software packages for GLMMs. With $\boldsymbol{\Sigma} = \tau\mathbf{D}$, i.e., defining the spatial correlation directly via the $\mathbf{D}$ matrix, the model can be fitted using earlier versions of the **hglm** package. However, for the general structure in Equations (3) and (4), estimation using explicit maximization of the marginal or profile likelihood involves quite advanced derivations as a consequence of partial differentiation of the likelihood including $\boldsymbol{\Sigma}$ with ρ and τ as parameters (see, e.g., Lee and Lee, 2012). In this paper, we show that by using eigen decomposition of $\boldsymbol{\Sigma}$, we can modify an already existing R package, e.g., **hglm**, with minor programming effort, to fit a HGLM with CAR or SAR random effects.

h-likelihood estimation

In order to explain the specific model fitting algorithm implemented in **hglm**, we present a brief overview of the EQL algorithm for HGLMs (Lee and Nelder, 1996). First, we start with the standard HGLM containing only uncorrelated random effects. Then, we extend the discussion for CAR and SAR random effects. Though it is possible to have more than one random-effect term, in a HGLM, coming from different distributions among the conjugate distributions to GLM families, for presentational simplicity we consider only one random-effect in this section. The (log-)h-likelihood for a HGLM with independent random effects can be presented as

$$h = \sum_i\sum_j\left(\frac{\left(y_{i,j}\theta_{i,j}-b\left(\theta_{i,j}\right)\right)}{\phi}+c\left(y_{i,j},\phi\right)\right)+\sum_i\left(\frac{\psi v_i - b_R\left(v_i\right)}{\tau}+c_R\left(\tau\right)\right), \tag{5}$$

where, $i=1,2,\ldots,n$, $j=1,2,\ldots k$, θ is the canonical parameter of the mean model, ϕ is the dispersion parameter of th mean model, $b\left(.\right)$ a function which satisfies $E\left(y_{i,j}|u_i\right) = \mu_{i,j} = b'\left(\theta_{i,j}\right)$ and $Var\left(y_{i,j}|u_i\right) = \phi b''\left(\theta_{i,j}\right) = \phi V\left(\mu_{i,j}\right)$, where $V(.)$ is the GLM variance function. Furthermore, ϕ is the dispersion parameter of the mean model, $\theta_R\left(u_i\right) = v_i$ with $\theta_R\left(.\right)$ being a so-called weak canonical link for u_i (leading to conjugate HGLMs) or an identity link for Gaussian random effects leading to GLMMs (Lee et al., 2006, pp. 3–4), $\psi = E\left(u_i\right)$, τ is the dispersion parameter, and b_R and c_R are some known functions depending on the distribution of u_i. Equation (5) also implies that h can be defined uniquely by using the means and the mean-variance relations of $y_{i,j}|u_i$ and the quasi-response ψ, allowing us representing it as a sum of two extended quasi-likelihoods (double EQL; Lee and Nelder, 2003) as

$$-2D = \sum_i\sum_j\left(\frac{d_{0,i,j}}{\phi}+\log\left(2\pi\phi V\left(y_{i,j}|u_i\right)\right)\right)+\sum_i\left(\frac{d_{1,i}}{\tau}+\log\left(2\pi\tau V_1\left(u_i\right)\right)\right), \tag{6}$$

where $d_{0,i,j}$ and $d_{1,i}$ are the deviance components of $y_{i,j}|u_i$ and ψ respectively and $Var(\psi) = \tau V_1(u_i)$. It is worth noting that $-2D$ is an approximation to h if $y|u$ or u (or both) belongs/belong to Binomial, Poisson, double-Poisson (or quasi-Poisson), Beta, or Gamma families. However, no such equivalent relation (even approximately) is known for the quasi-Binomial family.

In order to estimate the model parameters, Lee and Nelder (1996) suggested a two-step procedure in which the first, for fixed dispersion parameters ϕ and τ, h (or equivalently $-2D$) is maximized w.r.t. $\boldsymbol{\beta}$ and $\mathbf{v} = \{v_i\}$. This maximization leads to an iteratively weighted least-square (IWLS) algorithm which solves

$$\mathbf{T}_a' \blacksquare^{-1} \mathbf{T}_a \boldsymbol{\delta} = \mathbf{T}_a' \blacksquare^{-1} \mathbf{y}_a,$$

where $\mathbf{T}_a = \begin{pmatrix} \mathbf{X} & \mathbf{Z} \\ \mathbf{0} & \mathbf{I}_n \end{pmatrix}$, $\boldsymbol{\delta} = \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{v} \end{pmatrix}$, $\mathbf{y_a} = \begin{pmatrix} \mathbf{y}_{0,a} \\ \mathbf{y}_{1,a} \end{pmatrix}$ with $\mathbf{y}_{0,a} = \left\{ \eta_{i,j} - \left(y_{i,j} - \mu_{i,j} \right) \frac{\partial \eta_{i,j}}{\partial \mu_{i,j}} \right\}$ and $\mathbf{y}_{1,a} = \left\{ v_i - (\psi_i - u_i) \frac{\partial v_i}{\partial u_i} \right\}$ are the vectors of GLM working responses for $y_{i,j}|u_i$ and v_i respectively, and $\blacksquare = \mathcal{I}\mathbf{W}^{-1}$ with $\mathcal{I}$ being a diagonal matrix whose first n diagonal elements are all equal to ϕ and the remaining k diagonal elements are all equal to τ and $\mathbf{W} = \text{diag}(\mathbf{w}_0, \mathbf{w}_1)$ where $\mathbf{w}_0 = \left\{ \left(\frac{\partial \mu_{i,j}}{\partial \eta_{i,j}} \right)^2 \frac{1}{V(\mu_{i,j})} \right\}$ and $\mathbf{w}_1 = \left\{ \left(\frac{\partial u_i}{\partial v_i} \right)^2 \frac{1}{V_1(u_i)} \right\}$.

In the second step, the following profile likelihood is maximized to estimate the dispersion parameters.

$$p_{\boldsymbol{\beta},\mathbf{v}}(h) = \left(h - \frac{1}{2} \left| \frac{\partial^2 h}{\partial(\boldsymbol{\beta},\mathbf{v}) \partial(\boldsymbol{\beta},\mathbf{v})^T} \right| \right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}}, \tag{7}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are obtained from the fist step. One needs to iterate between the two steps until convergence. This procedure is often referred to as "HL(0,1)" (Lee and Lee, 2012) and is available in the R packages **spaMM** and **HGLMMM** (Molas and Lesaffre, 2011). Because there is no unified algorithm to maximize $p_{\boldsymbol{\beta},\mathbf{v}}$ computer packages often carry out the maximization by using general purpose optimization routines, e.g., package **spaMM** uses the `optim` function.

A unified algorithm for estimating the variance components can be derived by using profile likelihood adjustment in Equation (6) instead of (5). This leads to maximizing

$$p_{\boldsymbol{\beta},\mathbf{v}}(Q) = \left(-2D - \frac{1}{2} \left| \frac{\partial^2(-2D)}{\partial(\boldsymbol{\beta},\mathbf{v}) \partial(\boldsymbol{\beta},\mathbf{v})^T} \right| \right)_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}},\mathbf{v}=\hat{\mathbf{v}}} \tag{8}$$

for estimating τ and ϕ. The corresponding score equation of the dispersion parameters, after ignoring the fact that $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{v}}$ are functions of ϕ and τ, can be shown (see, e.g., Lee and Nelder, 2001) to have the form of Gamma family GLMs with $d_{0,i,j}/\left(1 - h_{i,j}\right)$ and $d_{1,i}/(1 - h_i)$ as the responses for ϕ and τ respectively, where $h_{i,j}$ and h_i are the hat values corresponding to $y_{i,j}$ and v_i in the first step, and $\left(1 - h_{i,j}\right)/2$ and $(1 - h_i)/2$ as the respective prior weights. This procedure is often referred to as EQL (Lee et al., 2006) or DEQL (Lee and Nelder, 2003) and was the only available procedure in **hglm** ($\leqslant$ 1.2-8). An advantage of this algorithm is that fixed effects can also be fitted in the dispersion parameters (Lee et al., 2006) without requiring any major change in the algorithm.

HL(0,1) and EQL were found to be biased, especially for binary responses when the cluster size is small (Lee and Nelder, 2001; Noh and Lee, 2007) and when τ is large in both Binomial and Poisson GLMMs. Several adjustments are suggested and explained in the literature (see, e.g., Lee and Nelder, 2001, 2003; Lee and Lee, 2012) to improve the performance of h-likelihood estimation. Among these alternative suggestions HL(1,1) is the most easily implementable and found to be computationally faster than the other alternatives (Lee and Lee, 2012). The HL(1,1) estimates the fixed effects, $\boldsymbol{\beta}$, by maximizing $p_{\mathbf{v}}(h)$ instead of h and can be implemented by adjusting the working response $\mathbf{y}_a$ in the IWLS step, according to Lee and Lee (2012). In **hglm** ($\geqslant$ 2.0), this correction for the $\boldsymbol{\beta}$ estimates has been implemented, through the `method = "EQL1"` option, but still $p_{\boldsymbol{\beta},\mathbf{v}}(Q)$ is maximized for the estimation of the dispersion parameters, in order to make use of the unified algorithm for dispersion parameter estimates via Gamma GLMs. Table 1 shows available options for h-likelihood estimation (subject to the corrections mentioned above) in different R packages.

The above EQL method cannot be directly applied to HGLMs with CAR/SAR random effects because the resulting $p_{\boldsymbol{\beta},\mathbf{v}}(Q)$ does not allow us to use the Gamma GLM to estimate the dispersion parameters, τ and ρ. In the following subsection we present a simplification which allows us to use the Gamma GLM for estimating τ and ρ.

	hglm	**spaMM**	**HGLMMM**
EQL	`method = "EQL"`	`HLmethod = "EQL-"`	-
HL(0,1)	-	`HLmethod = "HL(0,1)"`	`LapFix = FALSE`
HL(1,1)	[a]	`HLmethod = "HL(1,1)"`	`LapFix = TRUE`

[a] By specifying `method = "EQL1"` in the **hglm** package, the HL(1,1) correction of the working response is applied in the EQL algorithm.

Table 1: Implemented h-likelihood methods in **hglm** compared to **spaMM** and **HGLMMM** packages.

Simplification of model computation

The main difference between an ordinary GLMM with independent random effects and CAR/SAR random effects models is that the random effects in the later cases are not independent. In the following we show that by using the eigen decomposition we can reformulate a GLMM with CAR or SAR random effects to an equivalent GLMM with independent but heteroscedastic random effects.

Lemma 1 Let $\boldsymbol{\omega} = \{\omega_i\}_{i=1}^n$ be the eigenvalues of $\mathbf{D}$, and $\mathbf{V}$ is the matrix whose columns are the corresponding orthonormal eigenvectors, then

$$\frac{1}{\tau}(\mathbf{I} - \rho\mathbf{D}) = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T, \tag{9}$$

where $\boldsymbol{\Lambda}$ is diagonal matrix whose ith diagonal element is given by

$$\lambda_i = \frac{1 - \rho\omega_i}{\tau}. \tag{10}$$

Proof of Lemma 1

$$\begin{aligned} \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T &= \mathbf{V}\text{diag}\left\{\frac{1-\rho\omega_i}{\tau}\right\}\mathbf{V}^T \\ &= \frac{1}{\tau}(\mathbf{V}\mathbf{V}^T - \rho\mathbf{V}\text{diag}\{\omega_i\}\mathbf{V}^T) \\ &= \frac{1}{\tau}(\mathbf{I} - \rho\mathbf{D}) \end{aligned} \tag{11}$$

It is worth noting that the relation between the eigenvalues of $\mathbf{D}$ and $\boldsymbol{\Sigma}$ was already known as early as in Ord (1975) and was used to simplify the likelihood function of the Gaussian CAR model. Similarly, for simplification of the SAR model, we have

$$\frac{1}{\tau}(\mathbf{I} - \rho\mathbf{D})(\mathbf{I} - \rho\mathbf{D}^T) = \mathbf{V}\text{diag}\left\{\frac{(1-\rho\omega_i)^2}{\tau}\right\}\mathbf{V}^T. \tag{12}$$

An anonymous referee has pointed out that a somewhat extended version of **Lemma 1**, which deals with simultaneous diagonalization of two positive semi-definite matrices (Newcomb, 1969), has already been used to simplify the computation of Gaussian response models with intrinsic CAR random effects, especially in a Bayesian context (see, e.g., He et al., 2007, and the related discussions in Hodges 2013, Chap. 5). Therefore, **Lemma 1** is not an original contribution of this paper, however, to the authors' knowledge, no software package has yet utilized this convenient relationship to fit any HGLM by using interconnected GLMs, as discussed below.

Re-arranging Equation (1), we have

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta} + \tilde{\mathbf{Z}}\mathbf{u}^*, \tag{13}$$

where, $\boldsymbol{\eta} = \{\eta_s\}$, $\mathbf{X} = \{\mathbf{X}_s\}$, $\tilde{\mathbf{Z}} = \{\mathbf{Z}_s\}\,\mathbf{V}$ and $\mathbf{u}^* = \mathbf{V}^T\mathbf{u}$. With the virtue of **Lemma 1** and the properties of the multivariate normal distribution, we see that $\mathbf{u}^* \sim N\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right)$. This model can now easily be fitted using the **hglm** package in R. Further note that, for the CAR model, from **Lemma 1**, we have

$$\begin{aligned} \lambda_i &= \frac{1-\rho\omega_i}{\tau} \\ &= \theta_0 + \theta_1\omega_i \end{aligned} \tag{14}$$

Option	Explanation
`rand.family = CAR(D = nbr)`	The random effect has conditional or simul-
`rand.family = SAR(D = nbr)`	taneous autoregressive covariance structure. Here `nbr` is a matrix provided by the user.
`method = "EQL1"`	The first order correction of fixed effects (Lee and Lee, 2012) applied on the EQL estimates.
`rand.disp.X = X`	Linear predictor for the variance component of a random effect. The matrix `X` is provided by the user.
`rand.family = list(Gamma(),CAR())`	This option provides the possibility of having different distributions for the random effects.

Table 2: New options in the `hglm` function.

where $\theta_0 = 1/\tau$ and $\theta_1 = -\rho/\tau$. While for the SAR model,

$$\begin{aligned} \lambda_i &= \frac{(1-\rho\omega_i)^2}{\tau} \\ \Rightarrow \quad \sqrt{\lambda_i} &= \theta_0 + \theta_1\omega_i, \end{aligned} \tag{15}$$

where $\theta_0 = 1/\sqrt{\tau}$ and $\theta_1 = -\rho/\sqrt{\tau}$. Following Lee et al. (2006), we can use an inverse link for CAR or an inverse square-root link for SAR in a Gamma GLM with response ${u_i}^{*2}/\left(1-h_i\right)$, where h_i is the corresponding hat value from the mean model (Equation (13)), $\left(1-h_i\right)/2$ as the weight and the linear predictor given by Equations (14) and (15) to obtain an EQL estimate of θ_0 and θ_1.

Implementation

The spatial models above are implemented in the **hglm** package ($\geqslant$ 2.0) by defining new families, `CAR()` and `SAR()`, for the random effects. The "CAR" and "SAR" families allow the user to define spatial structures by specifying the **D** matrix. Using **Lemma 1**, these families are created using **D** with default link function "identity" for the Gaussian random effects **u**, and "inverse" and "inverse.sqrt" for the dispersion models (14) and (15). When the "CAR" or "SAR" family is specified for the random effects, the parameter estimates $\hat{\rho}$ and $\hat{\tau}$ are given in the output in addition to the other summary statistics of the dispersion models. The new input options in **hglm** are described in Table 2.

Throughout the examples below, we use Gaussian CAR random effects to introduce spatial correlation. However, one can also add additional independent non-Gaussian random effects along with the Gaussian CAR random effect. For example, an overdispersed count response can be fitted using a Poisson HGLM with an independent Gamma and Gaussian CAR random effects.

Examples and simulation study

In the **hglm** package vignette, we look at the improvement of EQL1 in comparison to EQL. Here, we focus on the precision of the parameter estimates with spatial HGLMs.

Poisson CAR & SAR model

We study the properties of the estimates produced by **hglm** using a simulation study built around the Scottish Lip Cancer example (see also the examples). We simulate data with the same **X** values, offset and neighborhood matrix as in the Scotthis Lip Cancer example data. We use the true values of the parameters, after Lee and Lee (2012), as $\left(\text{intercept}, \boldsymbol{\beta}_{fpp}, \tau, \rho\right) = (0.25, 0.35, 1.5, 0.1)$. The parameter estimates for 1000 Monte Carlo iterations are summarized in Table 3.

Both the EQL and the EQL1 correction are slightly biased for τ and ρ (Table 3) though the absolute amount of bias is small and may be negligible in practical applications. The EQL1 correction mainly improves the estimates of the intercept term. There were convergence problems for a small number of replicates, which was not surprising given the small number of observations ($n = 56$) and that the simulated value for the spatial autocorrelation parameter ρ connects the effects in the different Scottish districts rather weakly. Such convergence problems can be addressed by pre-specifying better starting values. For the converged estimates the bias was small and negligible.

Parameter	True value	Bias in the estimation methods			
		CAR		SAR	
		EQL†	EQL1 correction†	EQL†	EQL1 correction†
intercept	0.2500	0.0351*	0.0105	0.0541*	0.0307
β_{fpp}	0.3500	−0.0018*	−0.0004	−0.0034*	−0.0022
$1/\tau$	0.6667	0.0660*	0.0658*	n/a	n/a
$-\rho/\tau$	−0.0667	0.0118*	0.0119*	n/a	n/a
$1/\sqrt{\tau}$	0.8165	n/a	n/a	0.0393*	0.0370*
$-\rho/\sqrt{\tau}$	−0.0817	n/a	n/a	0.0037*	0.0049*
τ	1.5000	−0.0770*	−0.0760*	−0.0880*	−0.0860*
ρ	0.1000	−0.0247*	−0.0250*	−0.0097*	−0.0102*

* Significantly different from 0 at the 5% level.
† The estimates are the means from 1000 replicates.

Table 3: Average bias in parameter estimates in the simulation using the Scottish Lip Cancer example.

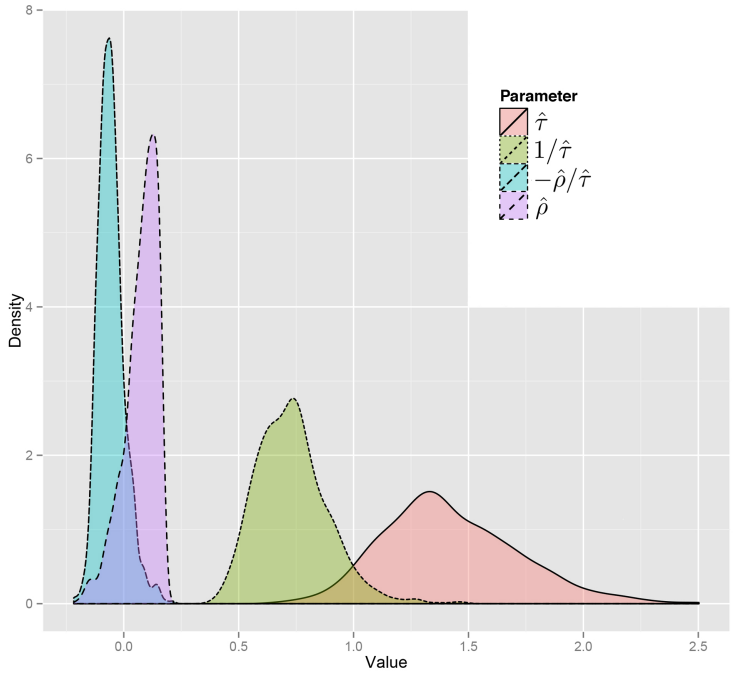

Figure 1: Density plot of the spatial variance-covariance parameter estimates from CAR models via `"EQL1"` over 1000 Monte Carlo simulations for the Scottish Lip Cancer example.

The simulation results also revealed that the distribution of $\hat{\rho}$ from CAR models is skewed (see Figure 1), which was also pointed out by Lee and Lee (2012). However, the distribution of $-\hat{\rho}/\hat{\tau}$ turned out to be less skewed than the distribution of $\hat{\rho}$. Similar observation was also found for SAR models. This suggests, we might draw any inference on spatial variance-covariance parameters in the transformed scale, θ_0 and θ_1.

Computational efficiency

Fitting CAR and SAR models for large data sets could be computationally challenging, especially for the spatial variance-covariance parameters. Using our new algorithm in **hglm**, moderately sized problems can be fitted efficiently.

We re-sampled from the `ohio` data set (for more details on the data set see also the examples) for different number of locations, each with 10 replicates, executed on a single Intel© Xeon© E5520 2.27GHz CPU. In each replicate, the data was fitted using both **hglm** and **spaMM**, for the same model described above. The results regarding average computational time are summarized in Figure 2. **hglm** is clearly more usable for fitting larger sized data sets. In the package vignette, we also show the comparisons of the parameter estimates, where the `"EQL"` estimates from **hglm** are almost identical

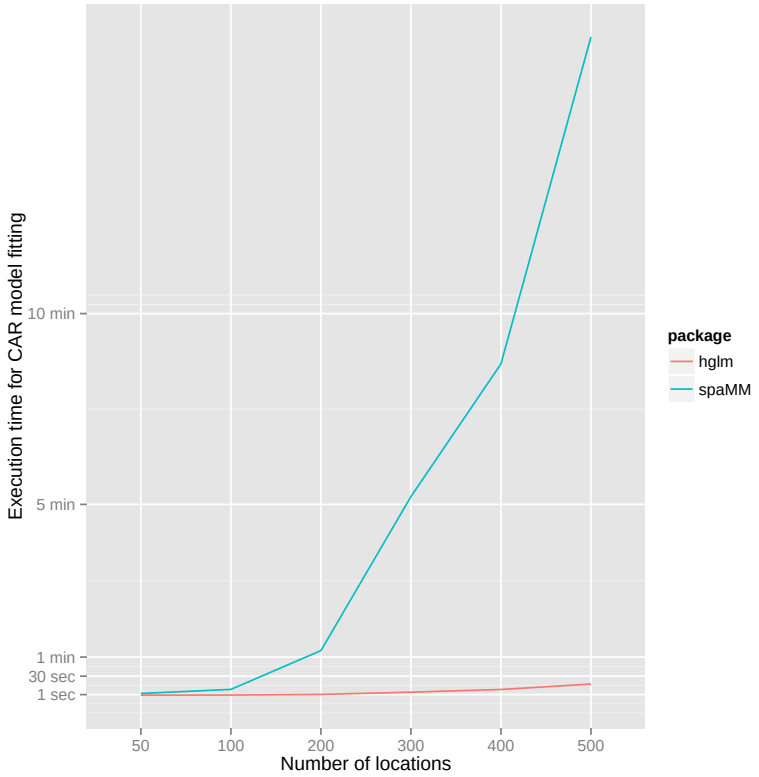

Figure 2: Comparison of the execution time for fitting CAR models using **hglm** and **spaMM**.

to the "EQL-" estimates from **spaMM**, with high correlation coefficients between the two methods: 1.0000, 0.9998, 0.9997 and 0.9999 for the residual variance, ρ, τ and the intercept, respectively.

Examples

Scottish Lip Cancer data set

Here we analyze the cancer data set (source: Clayton and Kaldor, 1987) from the **hglm** package. Calling data(cancer) loads a numeric vector E that represents the expected number of skin cancer patients in different districts in Scotland, a numeric vector O giving the corresponding observed counts, a numeric vector Paff giving proportion of people involved agriculture, farming, and fisheries, and matrix D giving the neighborhood structure among Scottish districts. Here we demonstrate how the data is fitted as a CAR model or a SAR model, using the **hglm** package with the EQL method.

```
> library(hglm)
> data(cancer)
> logE <- log(E)
> XX <- model.matrix(~ Paff)
> cancerCAR <- hglm(X = XX, y = O, Z = diag(56),
+                   family = poisson(),
+                   rand.family = CAR(D = nbr),
+                   offset = logE, conv = 1e-8,
+                   maxit = 200, fix.disp = 1)
> summary(cancerCAR)

Call:
hglm.default(X = XX, y = O, Z = diag(56), family = poisson(),
    rand.family = CAR(D = nbr), conv = 1e-08, fix.disp = 1, offset = logE)

----------
MEAN MODEL
----------
```

```
Summary of the fixed effects estimates:

            Estimate Std. Error t-value Pr(>|t|)
(Intercept)  0.26740    0.20732   1.290  0.20893
Paff         0.03771    0.01215   3.103  0.00471 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Note: P-values are based on 25 degrees of freedom

Summary of the random effects estimates:

     Estimate Std. Error
[1,]   0.6407     1.0467
[2,]   0.5533     0.3829
[3,]   0.4124     0.5202
...
NOTE: to show all the random effects, use print(summary(hglm.object),
      print.ranef = TRUE).

----------------
DISPERSION MODEL
----------------

NOTE: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 1

Model estimates for the dispersion term:

Link = log

Effects:
[1] 1

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

Dispersion parameter for the random effects:
[1] 656.3

Dispersion model for the random effects:

Link = log

Effects:
.|Random1
                 Estimate Std. Error
1/CAR.tau           6.487      1.727
-CAR.rho/CAR.tau   -1.129      0.303
CAR.tau (estimated spatial variance component): 0.1542
CAR.rho (estimated spatial correlation): 0.174

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

EQL estimation converged in 10 iterations.
```

In the above output provided by the `summary()` method, fixed effects estimates are given under `MEAN MODEL`, and the dispersion parameter estimates are given under `DISPERSION MODEL`. Here, we have only one random effects term that has a CAR structure, and the corresponding parameter estimates, $\hat{\theta}_0 = 6.487$ and $\hat{\theta}_1 = -1.129$, are given under `.|Random1`. However, these are not the natural parameters of the CAR model (see Section Simplification of model computation) therefore the estimates of the natural dispersion parameters are given just after them which are, in this case, $\hat{\tau} = 0.15$ and $\hat{\rho} = 0.17$.

Because the **D** matrix was a neighborhood matrix (consisting of 0's and 1's), the results imply that the partial correlation of the random effect for any two neighboring districts, given the same for all other districts, is 0.17.

Furthermore, the value 656.3 for the dispersion of the random effects given in the output above is an overall variance of $\mathbf{u}^*$ in Equation (13). This output is usually not of interest to the user and the main results are contained in $\hat{\tau}$ and $\hat{\rho}$.

```
> cancerSAR <- hglm(X = XX, y = O, Z = diag(56), family = poisson(),
+                   rand.family = SAR(D = nbr), offset = logE,
+                   conv = 1e-08, fix.disp = 1)
> summary(cancerSAR)

Call:
hglm.default(X = XX, y = O, Z = diag(56), family = poisson(),
    rand.family = SAR(D = nbr), conv = 1e-08, fix.disp = 1, offset = logE)

----------
MEAN MODEL
----------

Summary of the fixed effects estimates:

            Estimate Std. Error t-value Pr(>|t|)
(Intercept)  0.19579    0.20260   0.966  0.34241
Paff         0.03637    0.01165   3.122  0.00425 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Note: P-values are based on 27 degrees of freedom

Summary of the random effects estimates:

     Estimate Std. Error
[1,]   0.7367     1.0469
[2,]   0.6336     0.3930
[3,]   0.4537     0.5784
...
NOTE: to show all the random effects, use print(summary(hglm.object),
      print.ranef = TRUE).

----------------
DISPERSION MODEL
----------------

NOTE: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 1

Model estimates for the dispersion term:

Link = log

Effects:
[1] 1

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

Dispersion parameter for the random effects:
[1] 16.3

Dispersion model for the random effects:

Link = log
```

```
Effects:
.|Random1
                Estimate Std. Error
1/sqrt(SAR.tau)          2.7911     0.4058
-SAR.rho/sqrt(SAR.tau)  -0.4397     0.0822
SAR.tau (estimated spatial variance component): 0.1284
SAR.rho (estimated spatial correlation): 0.1575

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

EQL estimation converged in 12 iterations.
```

For the CAR model, the `hglm` estimates are exactly the same as those labelled as "PQL" estimates in Lee and Lee (2012). To get the EQL1 correction the user has to add the option `method = "EQL1"` and the `hglm` function gives similar results to those reported in Lee and Lee (2012), e.g., their HL(1,1) estimates were $\left(\text{intercept}, \boldsymbol{\beta}_{fpp}, \tau, \rho\right) = (0.238, 0.0376, 0.155, 0.174)$ whereas our `"EQL1"` correction gives (0.234, 0.0377, 0.156, 0.174) (see also Section "Fitting a spatial Markov Random Field model using the CAR family" in the package vignette). A minor difference to the `"EQL1"` result appears because Lee and Lee (2012) used the HL(1,1) modification to their HL(0,1) whereas we apply such a correction directly to EQL which is slightly different from HL(0,1) (see Table 1).

Ohio elementary school grades data set

We analyze a data set consisting of the student grades of 1,967 Ohio Elementary Schools during the year 2001–2002. The data set is freely available on the internet (URL `http://www.spatial-econometrics.com/`) as a web supplement to LeSage and Pace (2009) but was not analyzed therein. The shape files were downloaded from `http://www.census.gov/cgi-bin/geo/shapefiles2013/main` and the districts of 1,860 schools in these two files could be connected unambiguously. The data set contains information on, for instance, school building ID, Zip code of the location of the school, proportion of passing on five subjects, number of teachers, number of students, etc. We regress the median of 4th grade proficiency scores, **y**, on an intercept, based on school districts. The statistical model is given as

$$y_{i,j} = \mu + v_j + \epsilon_{i,j}, \tag{16}$$

where $i = 1, 2, \ldots, 1860$ (observations), $j = 1, 2, \ldots, 616$ (districts), $\epsilon_{i,j} \sim N\left(0, \sigma_e^2\right)$, $\{v_j\} = \mathbf{v} \sim N\left(\mathbf{0}, \tau\left(\mathbf{I} - \rho\mathbf{W}\right)^{-1}\right)$ and $\mathbf{W} = \{w_{p,q}\}_{p,q=1}^{616}$ is a spatial weight matrix (i.e., the neighborhood matrix). We construct $w_{p,q} = 1$ if the two districts p and q are adjacent, and $w_{p,q} = 0$ otherwise.

The above choice of constructing the weight matrix is rather simple. Because the aim of this paper is to demonstrate the use of **hglm** for fitting spatial models rather than drawing conclusions from real data analysis, we skip any further discussion on the construction of the weight matrix. Interested, readers are referred to LeSage and Pace (2009) for more discussion on the construction of the spatial weight matrices. With the spatial weight matrix defined as above, we can estimate model (16) by using our **hglm** package in the following way.

```
> ## load the data object 'ohio'
> data(ohio)
>
> ## fit a CAR model for the median scores of the districts
> X <- model.matrix(MedianScore ~ 1, data = ohioMedian)
> Z <- model.matrix(~ 0 + district, data = ohioMedian)
> ohioCAR <- hglm(y = ohioMedian$MedianScore, X = X, Z = Z,
+                 rand.family = CAR(D = ohioDistrictDistMat))
> summary(ohioCAR)

Call:
hglm.default(X = X, y = ohioMedian$MedianScore, Z = Z,
    rand.family = CAR(D = ohioDistrictDistMat))

----------
MEAN MODEL
----------
```

```
Summary of the fixed effects estimates:

            Estimate Std. Error t-value Pr(>|t|)
(Intercept)   72.429      0.819   88.44   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Note: P-values are based on 1566 degrees of freedom

Summary of the random effects estimates:

     Estimate Std. Error
[1,]  -21.433     11.071
[2,]  -17.890     10.511
[3,]   -4.537      7.844
...
NOTE: to show all the random effects, use print(summary(hglm.object),
      print.ranef = TRUE).

----------------
DISPERSION MODEL
----------------

NOTE: h-likelihood estimates through EQL can be biased.

Dispersion parameter for the mean model:
[1] 190.5

Model estimates for the dispersion term:

Link = log

Effects:
  Estimate Std. Error
    5.2498     0.0357

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

Dispersion parameter for the random effects:
[1] 1.01

Dispersion model for the random effects:

Link = log

Effects:
.|Random1
                Estimate Std. Error
1/CAR.tau         0.0097      8e-04
-CAR.rho/CAR.tau  -0.0011      2e-04
CAR.tau (estimated spatial variance component): 103.6
CAR.rho (estimated spatial correlation): 0.1089

Dispersion = 1 is used in Gamma model on deviances to calculate the
standard error(s).

EQL estimation converged in 5 iterations.
```

The estimated spatial correlation parameter among school districts is 0.109. We can obtain fitted values from the CAR model and predict the school districts without any observations. The following codes perform such prediction and the results are visualized in Figure 3(B). We remove the estimate of Lake Erie, as estimation for an uninhabited region is meaningless.

```
> ## extract districts from the map data
```

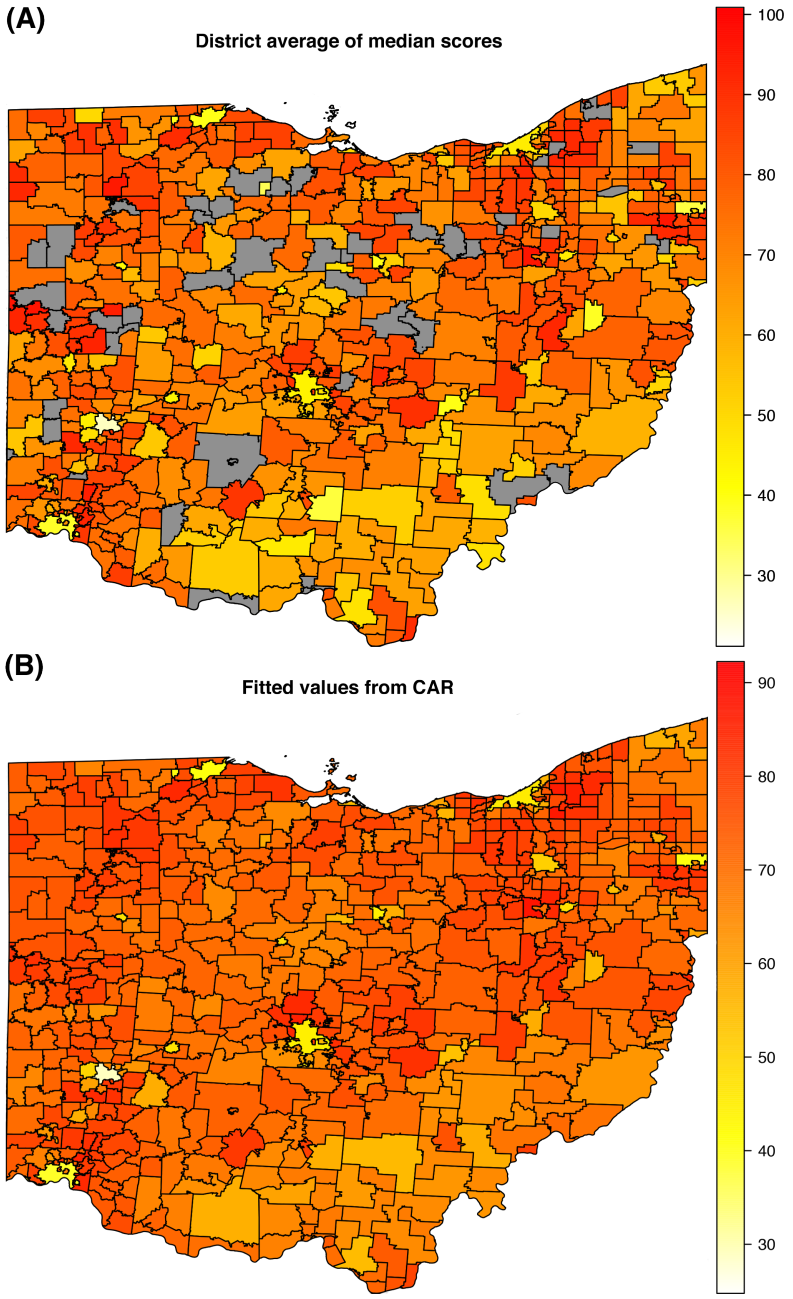

Figure 3: Observed (A) and predicted (B) median 4th grade proficiency scores of the school districts in Ohio. Districts without any observations are displayed in gray.

```
> districtShape <- as.numeric(substr(as.character(ohioShape@data$UNSDIDFP), 3, 7))
>
> ## calculate fitted values from the CAR model
> CARfit <- matrix(ohioCAR$ranef + ohioCAR$fixef,
+                  dimnames = list(rownames(ohioDistrictDistMat), NULL))
> ohioShape@data$CAR <- CARfit[as.character(districtShape),]
> is.na(ohioShape@data$CAR[353]) <- TRUE  # remove estimate of Lake Erie
>
> ## visualize the results
> spplot(ohioShape, zcol = "CAR", main = "Fitted values from CAR",
+        col.regions = heat.colors(1000)[1000:1], cuts = 1000)
```

A `predict()` method is not available because predicting spatially correlated random effects for autoregressive models requires re-fitting the whole model. Thus standard kriging cannot be used because the covariance structure changes if the neighborhood matrix is altered, while keeping τ and ρ

unchanged. Instead, the fitted model needs to include the entire neighborhood matrix with districts having missing data as well. Consequently, the incidence matrix $\boldsymbol{Z}$ has more columns than rows. This method of predicting random effects is frequently used in animal breeding applications (Henderson, 1984) but, to our knowledge, has not been applied to spatial autoregressive models previously.

In the example above, `ohioMedian$district` has 616 levels and 54 of the districts have no records (Figure 3(A)). The incidence matrix `Z`, created using the `model.matrix` function, therefore has 616 columns and 54 of these are columns of zeros. Hence, there are 616 levels in the fitted spatial random effect giving predictions for the districts without records.

Conclusion

The **hglm** package is one of few non-Bayesian packages on CRAN to fit spatial HGLMs, where the fixed and random effects are estimated simultaneously. We have shown how the HGLM framework, allowing linear predictors to model variance components, can be exploited to fit CAR and SAR models. This gives a computationally efficient algorithm for moderately sized problems (number of locations $<$ approx. 5000).

Acknowledgement

X. Shen was supported by a Swedish Research Council grant (No. 2014-371).

Bibliography

J. Besag. Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974. [p5]

D. Clayton and J. Kaldor. Empirical bayes estimation of age-standardized relative risk for use in disease mapping. *Biometrics*, 43:671–681, 1987. [p6, 11]

N. A. C. Cressie. *Statistics for Spatial Data*. Wiley, New York, revised edition, 1993. [p5, 6]

Y. He, J. S. Hodges, and B. P. Carlin. Re-considering the variance parameterization in multiple precision models. *Bayesian Analysis*, 2:529–556, 2007. [p8]

C. R. Henderson. *Applications of Linear Models in Animal Breeding*. University of Guelph, Guelph, Ontario, 1984. [p17]

J. S. Hodges. *Richly Parametarized Linear Models: Additive Linear and Time Series and Spatial Models Using Random Effects*. Chapman and Hall/CRC, Boca Raton, 2013. [p6, 8]

W. Lee and Y. Lee. Modifications of REML algorithm for HGLMs. *Statistics and Computing*, 22:959–966, 2012. [p5, 6, 7, 9, 10, 14]

Y. Lee and J. A. Nelder. Hierarchical generalized linear models (with discussion). *Journal of the Royal Statistical Society, Series B*, 58:619–678, 1996. [p5, 6, 7]

Y. Lee and J. A. Nelder. Hierarchical generalised linear models: A synthesis of generalised linear models, random effect models and structured dispersions. *Biometrika*, 88:987–1006, 2001. [p7]

Y. Lee and J. A. Nelder. Extended-REML estimators. *Journal of Applied Statistics*, 30:845–846, 2003. [p6, 7]

Y. Lee, J. A. Nelder, and Y. Pawitan. *Generalized Linear Models with Random Effects: Unified Analysis via H-Likelihood*. Chapman & Hall/CRC, Boca Raton, 2006. [p5, 6, 7, 9]

J. LeSage and R. Pace. *Introduction to Spatial Econometrics*. Chapman & Hall/CRC, Boca Raton, 2009. [p14]

F. Lindgren and H. Rue. Bayesian spatial modelling with R-INLA. *Journal of Statistical Software*, 63(19): 1–25, 2015. URL http://www.jstatsoft.org/v63/i19. [p5]

D. J. Lunn, J. A. Thomas, N. Best, and D. Spiegelhalter. WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, 10:325–337, 2000. [p5]

T. G. Martins, D. Simpson, F. Lindgren, and H. Rue. Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67:68–83, 2013. [p5]

M. Molas and E. Lesaffre. Hierarchical generalized linear models: The R package HGLMMM. *Journal of Statistical Software*, 39(13):1–20, 2011. URL http://www.jstatsoft.org/v39/i13. [p7]

R. Newcomb. On the simultaneous diagonalization of two semi-definite matrices. *Quarterly of Applied Mathematics*, 19:144–146, 1969. [p8]

M. Noh and Y. Lee. REML estimation for binary data in GLMMs. *Journal of Multivariate Analysis*, 98: 896–915, 2007. [p5, 7]

K. Ord. Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70:120–126, 1975. [p8]

L. Rönnegård, X. Shen, and M. Alam. hglm: A package for fitting hierarchical generalized linear models. *The R Journal*, 2(2):20–28, 2010. [p5]

F. Rousset and J.-B. Ferdy. Testing environmental and genetic effects in the presence of spatial autocorrelation. *Ecography*, 37(8):781–790, 2014. [p5]

M. M. Wall. A close look at the spatial structure implied by the CAR and SAR models. *Journal of Statistical Planning and Inference*, 121:311–324, 2004. [p6]

Moudud Alam
Statistics, School of Technology and Business Studies
Dalarna University, Sweden
maa@du.se

Lars Rönnegård
Statistics, School of Technology and Business Studies
Dalarna University, Sweden
and
Department of Animal Breeding and Genetics
Swedish University of Agricultural Sciences, Sweden
and
Division of Computational Genetics
Department of Clinical Sciences
Swedish University of Agricultural Sciences, Sweden
lrn@du.se

Xia Shen
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet, Sweden
and
MRC Human Genetics Unit
MRC Institute of Genetics and Molecular Medicine
University of Edinburgh, United Kingdom
xia.shen@ki.se