Journal of the Royal Statistical Society
SERIES B Statistical Methodology

J. R. Statist. Soc. B (2013)
75, *Part* 1, *pp.* 139–159

Dimension reduction and alleviation of confounding for spatial generalized linear mixed models

John Hughes

University of Minnesota, Minneapolis, USA

and Murali Haran

Pennsylvania State University, University Park, USA

[Received November 2010. Final revision April 2012]

Summary. Non-Gaussian spatial data are very common in many disciplines. For instance, count data are common in disease mapping, and binary data are common in ecology. When fitting spatial regressions for such data, one needs to account for dependence to ensure reliable inference for the regression coefficients. The spatial generalized linear mixed model offers a very popular and flexible approach to modelling such data, but this model suffers from two major shortcomings: variance inflation due to spatial confounding and high dimensional spatial random effects that make fully Bayesian inference for such models computationally challenging. We propose a new parameterization of the spatial generalized linear mixed model that alleviates spatial confounding and speeds computation by greatly reducing the dimension of the spatial random effects. We illustrate the application of our approach to simulated binary, count and Gaussian spatial data sets, and to a large infant mortality data set.

Keywords: Dimension reduction; Generalized linear model; Harmonic analysis; Mixed model; Regression; Spatial statistics

1. Introduction

Over 20 years ago Besag *et al.* (1991) proposed the spatial generalized linear mixed model (SGLMM) for areal data, which is a hierarchical model that introduces spatial dependence through a latent Gaussian Markov random field (GMRF). Although Besag *et al.* (1991) focused only on prediction for count data, the SGLMM for areal data has since been applied to other types of data (e.g. binary), in many fields (e.g. ecology, geology and forestry), and for regression as well as prediction. In fact, the SGLMM has long been the dominant areal model, owing to its flexible hierarchical specification, to the availability of the WinBUGS software application (Lunn *et al.*, 2000), which greatly eases data analysis, and to the various theoretical and computational difficulties that plague the areal SGLMM's nearest competitor: the automodel (see, for example, Hughes *et al.* (2010) and Kaiser and Cressie (1997, 2000)).

SGLMMs, although remarkably flexible and widely applicable, suffer from two major shortcomings:

(a) variance inflation due to spatial confounding and
(b) computational challenges posed by high dimensional latent variables.

The issue of spatial confounding was discovered by Clayton *et al.* (1993) and recently reiterated

Address for correspondence: John Hughes, Division of Biostatistics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: hughes@umn.edu

 1369–7412/13/75139

by Reich *et al.* (2006), who suggested a new model (henceforth the RHZ model) that seeks to alleviate confounding by explicitly introducing synthetic predictors that are orthogonal to the fixed effects predictors. But the RHZ model, by failing to account for the underlying graph, permits structure in the random effects that corresponds to negative spatial dependence, i.e. repulsion, which we do not expect to observe in the phenomena to which these models are typically applied.

We propose a new model that mitigates confounding while including patterns of positive spatial dependence, i.e. attraction, in the random effects and excluding patterns of repulsion. We achieve this by exploiting what we believe to be the intrinsic geometry for these models. The utilization of this geometry permits a natural and dramatic reduction in the dimension of the random effects, which speeds computation to the extent that our model makes feasible the analyses of large areal data sets.

Several recent references have focused on dimension reduction for point level, i.e. Gaussian process-based, spatial models (see, for example, Higdon (2002), Cressie and Johannesson (2008), Banerjee *et al.* (2008), Furrer *et al.* (2006) and Rue and Tjelmeland (2002)). To our knowledge, this paper is the first to propose a principled approach to dimension reduction for areal models (while also improving regression inference).

The rest of the paper is organized as follows. In Section 2 we review the traditional SGLMM. In Section 3 we discuss spatial confounding and review the model that was proposed by Reich *et al.* (2006). In Section 4 we present our sparse reparameterization of the areal SGLMM. In Section 5 we discuss dimension reduction for spatial models. In Section 6 we discuss the results of applying each of the three models to simulated binary, count and Gaussian spatial data. In Section 7 we apply our sparse model to a large US infant mortality data set. We conclude with a discussion.

2. Traditional spatial generalized linear mixed model

Besag *et al.* (1991) formulated their SGLMM for areal data as follows. Let $G = (V, E)$ be the underlying undirected graph, where $V = \{1, 2, \ldots, n\}$ are the vertices (each of which represents an area of interest) and E are the edges (each of which is a pair, (i, j), that represents the proximity of areas i and j). In what follows we shall represent G by using its adjacency matrix $\mathbf{A}$, which is the $n \times n$ matrix with entries given by $\mathrm{diag}(\mathbf{A}) = \mathbf{0}$ and $\mathbf{A}_{ij} = \mathbf{1}\{(i, j) \in E, i \neq j\}$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function.

Now, let $\mathbf{Z} = (Z_1, \ldots, Z_n)'$ be the random field of interest, where Z_i is the random variable associated with vertex i. Then the first stage of the model is given by

$$g\{\mathbb{E}(Z_i|\boldsymbol{\beta}, W_i)\} = \mathbf{X}_i\boldsymbol{\beta} + W_i, \tag{1}$$

where g is a link function, $\mathbf{X}_i$ is the ith row of the design matrix, $\boldsymbol{\beta}$ is a p-vector of regression parameters and W_i is a spatial random effect associated with vertex i. As for classical generalized linear models, different types of data imply different (canonical) choices of the link function g. For example, the canonical link function for binary spatial data is the logit function, $\mathrm{logit}(p) = \log\{p/(1-p)\}$, for count data it is the natural logarithm function, and for normal data it is the identity function (Nelder and Wedderburn, 1972). In the last case we have the spatial linear mixed model, which is of considerable interest (see, for example, Cressie (1993) and Banerjee *et al.* (2004)).

The field of random effects, $\mathbf{W} = (W_1, \ldots, W_n)'$, whereby the traditional model incorporates spatial dependence, is assumed to follow the so-called intrinsic conditionally auto-regressive or GMRF prior:

$$p(\mathbf{W}|\tau) \propto \tau^{\text{rank}(\mathbf{Q})/2} \exp\left(-\frac{\tau}{2} \mathbf{W}'\mathbf{Q}\mathbf{W} \right), \tag{2}$$

where τ is a smoothing parameter and $\mathbf{Q} = \text{diag}(\mathbf{A}\mathbf{1}) - \mathbf{A}$ is a precision matrix ($\mathbf{1}$ is the conformable vector of 1s). Note that $\mathbf{Q}$ intuitively incorporates both dependences (W_i and W_j, $i \neq j$, are independent given their neighbours if and only if $\mathbf{Q}_{ij} = \mathbf{Q}_{ji} = 0$ if and only if $(i, j) \notin E$) and prior uncertainty (our uncertainty about W_i is inversely proportional to the degree of vertex i: $\mathbf{Q}_{ii} = \mathbf{A}_i\mathbf{1}$, where $\mathbf{A}_i$ denotes the ith row of $\mathbf{A}$).

As described in the seminal paper by Diggle *et al.* (1998), an SGLMM for spatial data over a continuous domain, i.e. for point level data, can be formulated by replacing the GMRF specification with that of a Gaussian process (also see De Oliveira (2000) and Christensen and Waagepetersen (2002)). See Banerjee *et al.* (2004), Rue and Held (2005) or Haran (2012) for an introduction to the use of Gaussian random field models in spatial statistics.

Since prior (2) is improper ($\mathbf{Q}$ is singular) the traditional areal model restricts us to a Bayesian or restricted maximum likelihood analysis, although it is possible to use a maximum likelihood approach if a proper GMRF distribution is used instead of the intrinsic GMRF (see Besag and Kooperberg (1995)). The models that are described in Sections 3 and 4 have conditional autoregressions with invertible precision matrices, and so those models lend themselves to both classical and Bayesian analyses. In the interest of concision we use only Bayesian terminology in what follows.

3. Spatial confounding

We begin with a discussion of confounding for the areal SGLMM's closest competitor, the automodel, because we believe that this offers insight regarding confounding for the SGLMM. The automodel, a Markov random-field model that incorporates dependence directly rather than hierarchically, can be specified as

$$g\{\mathbb{E}(Z_i|\boldsymbol{\beta}, \eta, \mathbf{Z}_{-i})\} = \mathbf{X}_i\boldsymbol{\beta} + \eta \sum_{(i,j)\in E} Z_j^*, \tag{3}$$

where g (here and in the remainder of the paper) is a canonical link function, $\mathbf{Z}_{-i}$ denotes the field with the ith observation excluded and η is the dependence parameter ($\eta > 0$ implies an attractive model; $\eta < 0$ a repulsive model). For the so-called uncentred automodel, $Z_j^* = Z_j$, whereas $Z_j^* = Z_j - \mu_j$ for the centred automodel, where μ_j is the independence expectation of Z_j, i.e. $\mu_j = \mathbb{E}(Z_j|\boldsymbol{\beta}, \eta = 0) = g^{-1}(\mathbf{X}_j\boldsymbol{\beta})$.

The sum in equation (3) is called the autocovariate. We see that it is a synthetic predictor that employs only the observations themselves or the observations along with the posited regression component $\mathbf{X}\boldsymbol{\beta}$ of the model. The centred form of the autocovariate leads easily to an interpretation of the dependence parameter: η measures the 'reactivity' of an observation to its neighbours, conditional on the hypothesized regression component. This interpretation of η lends $\boldsymbol{\beta}$ their desired interpretation as regression parameters. Put another way, the purpose of the centred autocovariate is to fit small-scale structure in the data, by which we mean structure that is residual to the large-scale structure that is represented by $\mathbf{X}\boldsymbol{\beta}$. Evidently the centred autocovariate is well suited to this role, since η tends to be at most weakly correlated with $\boldsymbol{\beta}$.

The uncentred automodel, in contrast, exhibits both conceptual and spatial confounding. It is not clear how we should interpret the uncentred autocovariate, and so η and $\boldsymbol{\beta}$ are also difficult to interpret. Moreover, η and $\boldsymbol{\beta}$ tend to be strongly correlated for the uncentred model. (See Caragea and Kaiser (2009) for a treatment of confounding in the context of the autologistic model.)

Reich *et al*. (2006) employed a reparameterization to show that the traditional SGLMM, like the uncentred automodel, is confounded. More specifically, they showed that introduction of the random effects can inflate the variance of the posterior distribution of $\boldsymbol{\beta}$. This is because the traditional model implicitly contains predictors that are collinear with $\mathbf{X}$, and it is this collinearity that causes the variance inflation. To see this, let $\mathbf{P}$ be the orthogonal projection onto $C(\mathbf{X})$, i.e.

$$\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

and let $\mathbf{P}^{\perp}$ be the projection onto $C(\mathbf{X})$'s orthogonal complement: $\mathbf{P}^{\perp} = \mathbf{I} - \mathbf{P}$. Now, spectrally decompose these operators to acquire orthogonal bases, $\mathbf{K}_{n\times p}$ and $\mathbf{L}_{n\times(n-p)}$, for $C(\mathbf{X})$ and $C(\mathbf{X})^{\perp}$ respectively. These bases allow us to rewrite equation (1) as

$$g\{\mathbb{E}(Z_i|\boldsymbol{\beta}, W_i)\} = \mathbf{X}_i\boldsymbol{\beta} + W_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{K}_i\boldsymbol{\gamma} + \mathbf{L}_i\boldsymbol{\delta},$$

where $\boldsymbol{\gamma}$ and $\boldsymbol{\delta}$ are random coefficients. This form exposes the source of the spatial confounding: $\mathbf{K}$ and $\mathbf{X}$ have the same column space.

Since the offending predictors $\mathbf{K}$ have no scientific meaning, Reich *et al*. (2006) suggested that they be deleted from the model. Setting $\boldsymbol{\gamma} = \mathbf{0}$ leads to the following specification. For the first stage we have

$$g\{\mathbb{E}(Z_i|\boldsymbol{\beta}, \boldsymbol{\delta})\} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{L}_i\boldsymbol{\delta}.$$

And the prior for the random effects, $\boldsymbol{\delta}$, is now

$$p(\boldsymbol{\delta}|\tau) \propto \tau^{(n-p)/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\delta}'\mathbf{Q}_{\mathrm{R}}\boldsymbol{\delta}\right),$$

where $\mathbf{Q}_{\mathrm{R}} = \mathbf{L}'\mathbf{Q}\mathbf{L}$.

Reich *et al*. (2006) referred to this as smoothing orthogonal to the fixed effects, and they showed that their model does, in principle, address the confounding of the traditional model. More specifically, adding this restricted conditional auto-regression to the non-spatial model adjusts, but does not inflate, the variance of $\boldsymbol{\beta}$'s posterior. We also note that Reich *et al*. (2006) reduced slightly the number of model parameters, from $n + p + 1$ to $n + 1$.

The traditional SGLMM and RHZ model take essentially the same approach as the automodel in the sense that all these models augment the linear predictor with synthetic predictors. The traditional model can be considered to employ implicitly the synthetic predictors $\mathbf{K}$ and $\mathbf{L}$, and the RHZ model explicitly employs $\mathbf{L}$. The traditional SGLMM is analogous to the uncentred automodel in that both models introduce predictors—$\mathbf{K}$ and the uncentred autocovariate, respectively—that cause spatial confounding and thus render $\boldsymbol{\beta}$ uninterpretable. And the RHZ model is analogous to the centred automodel in that both models introduce predictors that are designed to fit only residual structure in the data.

4. Sparse reparameterization of the areal spatial generalized linear mixed model

Although the synthetic predictors for the RHZ model are readily interpreted, receiving their interpretation from the theory of linear models, they do not permit parsimonious fitting of the residual clustering that arises because of spatial dependence. This is because the geometry corresponding to the operator $\mathbf{P}^{\perp}$ neglects the underlying graph G. In this section we reparameterize the areal SGLMM by using an alternative operator that captures the intrinsic geometry of these models. The random effects for our new model

(a) include patterns corresponding to positive spatial dependence only, i.e. repulsive patterns are excluded, and
(b) have dimension much smaller than n.

In an attempt to reveal the structure of missing spatial covariates, Griffith (2003) augmented a generalized linear model with selected eigenvectors of $(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{A}(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)$, where $\mathbf{I}$ is the $n \times n$ identity matrix and $\mathbf{1}$ is the n-vector of 1s. This operator appears in the numerator of Moran's I-statistic,

$$I(\mathbf{A}) = \frac{n}{\mathbf{1}'\mathbf{A}\mathbf{1}} \frac{\mathbf{Z}'(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{A}(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{Z}}{\mathbf{Z}'(\mathbf{I}-\mathbf{1}\mathbf{1}'/n)\mathbf{Z}},$$

which is a popular non-parametric measure of spatial dependence (Moran, 1950). We see that $I(\mathbf{A})$ is a scaled ratio of quadratic forms in $\mathbf{e} = \mathbf{Z} - \bar{Z}\mathbf{1}$. The numerator of the ratio is the squared length of $\mathbf{e}$ in the elliptical space corresponding to $\mathbf{A}$, whereas the denominator is the squared length of $\mathbf{e}$ in a spherical space.

Since we are interested not in missing covariates but in smoothing orthogonal to $\mathbf{X}$ (which may or may not contain $\mathbf{1}$), we replace $\mathbf{I}-\mathbf{1}\mathbf{1}'/n$ with $\mathbf{P}^{\perp}$. The resulting operator, $\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$, which we call the Moran operator for $\mathbf{X}$ with respect to G, appears in the numerator of a generalized form of Moran's I:

$$I_{\mathbf{X}}(\mathbf{A}) = \frac{n}{\mathbf{1}'\mathbf{A}\mathbf{1}} \frac{\mathbf{Z}'\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}\mathbf{Z}}{\mathbf{Z}'\mathbf{P}^{\perp}\mathbf{Z}}.$$

Boots and Tiefelsdorf (2000) showed that

(a) the (standardized) spectrum of a Moran operator comprises the possible values for the corresponding $I_{\mathbf{X}}(\mathbf{A})$ and
(b) the eigenvectors comprise all possible mutually distinct patterns of clustering residual to $\mathbf{X}$ and accounting for G.

The positive and negative eigenvalues correspond respectively to varying degrees of positive and negative spatial dependence, and the eigenvectors associated with a given eigenvalue (λ_i, say) are the patterns of spatial clustering that data exhibit when the dependence between them is of degree λ_i.

In the language of harmonic analysis (Katznelson, 2004), one might say that the residual (to $\mathbf{X}$) of a data set on G is a noisy superposition of basic waves, or harmonics. The harmonics are the eigenvectors of the Moran operator. The harmonics that correspond to positive frequencies (eigenvalues) are those harmonics that could arise under positive spatial dependence, whereas the harmonics that correspond to negative or zero frequencies are those that could arise under negative spatial dependence or independence respectively. Since we expect the typical areal data set to have arisen under positive spatial dependence, the residual of such a data set must be a superposition of positive harmonics only, in which case the residual has an entirely consonant 'sound'. Residual spatial repulsion, in contrast, would have a dissonant sound.

To illustrate the appropriateness of the Moran basis for parsimoniously fitting residual spatial clustering, we consider the 30×30 square lattice and $\mathbf{X} = [\mathbf{x} \quad \mathbf{y}]$, where the ith row of the design matrix contains the x- and y-co-ordinates of the ith lattice point, with co-ordinates restricted to the unit square. The panels of Fig. 1 show eigenvectors 7, 13 and 42 of the RHZ model basis $(\mathbf{L}_7, \mathbf{L}_{13}, \mathbf{L}_{42})$ and of the Moran basis $(\mathbf{M}_7, \mathbf{M}_{13}, \mathbf{M}_{42})$. Each panel displays its eigenvector as a 'map' by associating the ith component of the vector with the spatial location (x_i, y_i) of the ith lattice point. We note that $\mathbf{M}_7$, $\mathbf{M}_{13}$ and $\mathbf{M}_{42}$ are associated with eigenvalues 0.995, 0.970 and

0.868 respectively, which implies that these eigenvectors correspond to strong positive spatial dependence.

We chose eigenvectors 7, 13 and 42 at random. For the RHZ model, these three eigenvectors are as good as any other three because all eigenvectors of $\mathbf{P}^{\perp}$ look essentially the same. Each vector of the RHZ basis is relatively flat, with the flatness occasionally broken by a very localized spike or dip. This blandness of the RHZ basis is reflected in the spectrum of $\mathbf{P}^{\perp}$, which is not useful for selecting one eigenvector as being more relevant than another—since $\mathbf{P}^{\perp}$ is a projection, all eigenvalues are 0 or 1. The chosen vectors of the Moran basis, in contrast, show smooth patterns of spatial variation at various scales. By contrast, Moran eigenvectors that correspond to negative eigenvalues exhibit rough patterns of spatial variation. In either case, the magnitude of the eigenvalue that is associated with a given eigenvector indicates the scale of the eigenvector's spatial pattern.

We henceforth assume that the matrix $\mathbf{M}$ contains the first $q \ll n$ eigenvectors of the Moran operator. Since the spatial analyst typically expects neighbouring observations to be similar, only the smooth harmonics are of interest here, and so we assume that $\lambda_q > 0$. (This assumption may not be appropriate in other settings, e.g. where one might expect to see residual roughening. We return to this issue in the final section of the paper.) This means that we can discard at least half of the eigenvectors in many cases, and our simulated examples will show that a much greater reduction is often possible.

Replacing $\mathbf{L}$ with $\mathbf{M}$ in the RHZ model gives, for the first stage,

$$g\{\mathbb{E}(Z_i|\boldsymbol{\beta}, \boldsymbol{\delta}_S)\} = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{M}_i\boldsymbol{\delta}_S.$$

And the prior for the random effects is now

$$p(\boldsymbol{\delta}_S|\tau) \propto \tau^{q/2} \exp\left(-\frac{\tau}{2}\boldsymbol{\delta}_S'\mathbf{Q}_S\boldsymbol{\delta}_S \right),$$

where $\mathbf{Q}_S = \mathbf{M}'\mathbf{Q}\mathbf{M}$. This implies $p + q + 1$ model parameters.

It would be difficult, perhaps impossible, to provide an intuitive conditional interpretation of $\mathbf{Q}_S$ (or $\mathbf{Q}_R$), and we feel that such an interpretation would probably be misleading regarding the marginal dependence structure. This is because the typical (*a priori*) interpretations provided for the traditional conditionally auto-regressive models are conditional interpretations that are known to be misleading. The assumption is that those models imply positive spatial dependence between any pair of observations and that the dependence behaves in an intuitive fashion. But Wall (2004) showed that this is not so, and Assunção and Krainski (2009) recently showed why it is not so: the *a priori* covariance structure of the proper conditional auto-regression and the *a posteriori* covariance structure of the intrinsic conditional auto-regression both depend, in a rather complicated way, on the overall structure of G. Hence, we feel that relying on *a priori* intuition is misguided for these models, and so we instead focus on the dependence structure of the posterior distribution, which can be interpreted as residual structure in the spatial field.

This sparse model is more closely analogous to the centred automodel than is the RHZ model. The uncentred automodel accounts for the underlying graph, since the uncentred autocovariate is a sum over neighbours, but does not fit structure residual to $\mathbf{X}$. The RHZ model accounts for $\mathbf{X}$ but not for the underlying graph. Both the centred automodel and our sparse SGLMM account for $\mathbf{X}$ and for the underlying graph. Table 1 compares and contrasts the five models that are treated here.

A referee pointed out that fitting the RHZ and sparse models as specified above may not yield *a posteriori* uncorrelated fixed and random effects unless we adjust for the link function. Although it is true that a non-linear link function does induce some collinearity between $\mathbf{X}$ and

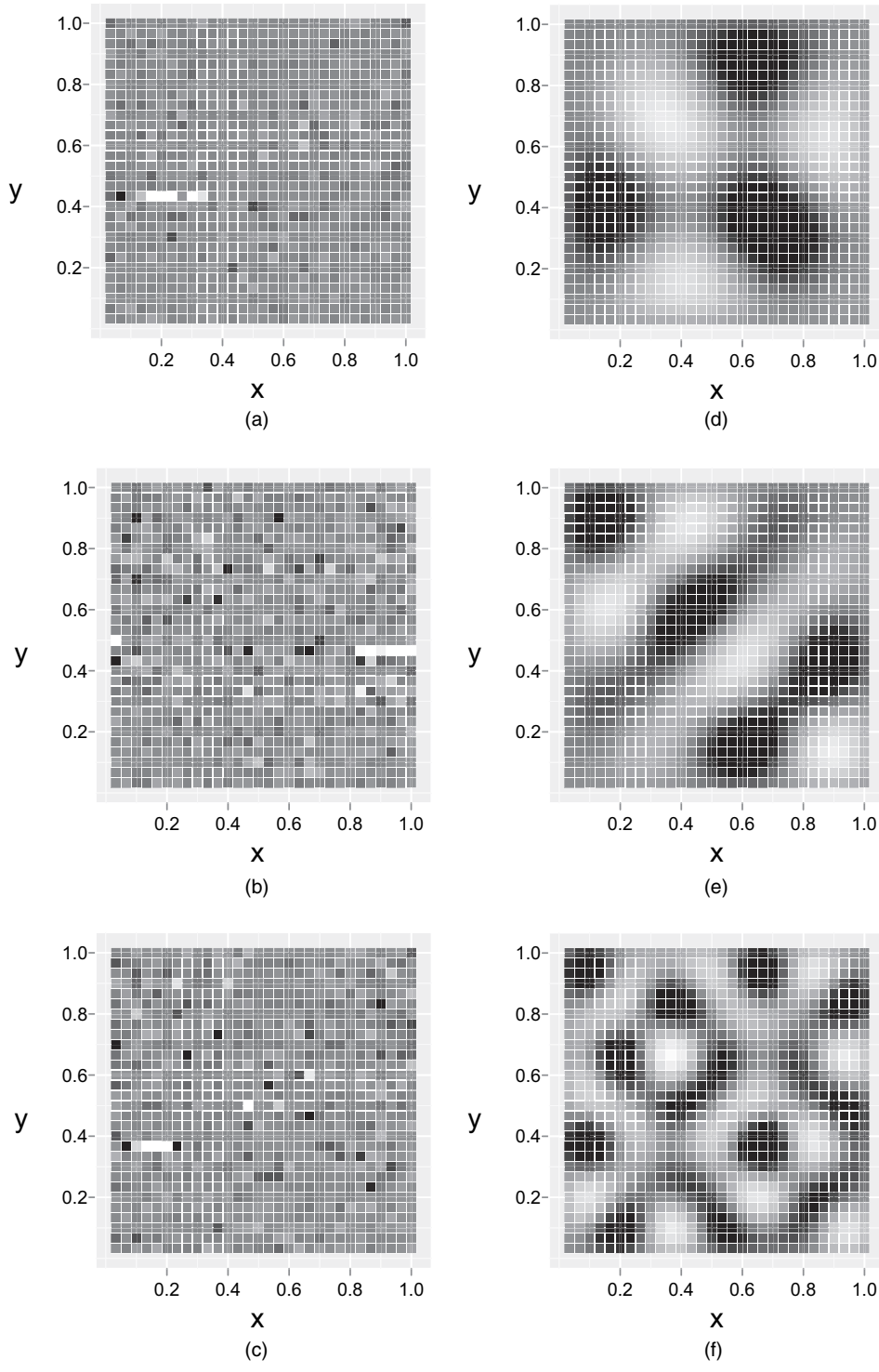

Fig. 1. Eigenvectors 7, 13 and 42 from the RHZ and Moran bases (the Moran basis appears to be an appropriate basis for parsimoniously fitting small-scale structure): (a) $\mathbf{L}_7$; (b) $\mathbf{L}_{13}$; (c) $\mathbf{L}_{42}$; (d) $\mathbf{M}_7$; (e) $\mathbf{M}_{13}$; (f) $\mathbf{M}_{42}$

Table 1. Comparing and contrasting various spatial generalized linear models

Model	*Systematic distortion of mean*	*Confounded*	*Accounts for G*
Uncentred automodel	$\eta \sum_{(i,j)\in E} Z_j$	Yes	Yes
Traditional SGLMM	$\mathbf{K}_i\boldsymbol{\gamma}+\mathbf{L}_i\boldsymbol{\delta}$	Yes	No
RHZ SGLMM	$\mathbf{L}_i\boldsymbol{\delta}$	No	No
Centred automodel	$\eta \sum_{(i,j)\in E}(Z_j-\mu_j)$	No	Yes
Sparse SGLMM	$\mathbf{M}_i\boldsymbol{\delta}_S$	No	Yes

L (or **M**), we found $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ (or $\boldsymbol{\delta}_S$) to be approximately *a posteriori* uncorrelated for every one of our more than 100 simulated data sets.

We note that another popular measure of spatial dependence is Geary's C, which uses the graph Laplacian **Q** in place of **A** (Geary, 1954). Geary's C is the spatial analogue of the well-known Durbin–Watson statistic for measuring auto-correlation in the residuals from a time series regression (Durbin and Watson, 1950). More specifically, the Durbin–Watson statistic is similar to Geary's C for a path graph (where adjacency is in time rather than space). Moran's I corresponds to a product moment formulation, and Geary's C to a squared difference formulation. The eigensystem of $\mathbf{P}^{\perp}\mathbf{Q}\mathbf{P}^{\perp}$ is a viable alternative to that of $\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$ for the current application, and perhaps other matrix representations of G would also provide suitable eigensystems. We chose **A** because the spectrum of a Moran operator is particularly easy to interpret.

5. Dimension reduction for spatial models

Fitting a Gaussian process model like that mentioned in Section 2 requires the repeated evaluation of expressions involving the inverse of the covariance matrix $\mathbf{H}(\phi)$, where ϕ are the parameters of the spatial covariance function. The customary approach to this problem is to avoid inversion in favour of Cholesky decomposition of **H** followed by a linear solve. Since **H** is typically dense, its Cholesky decomposition is in $O(n^3)$, and so the time complexity of the overall fitting algorithm is in $O(n^3)$. This considerable computational expense makes the analyses of large point level data sets time consuming or infeasible. Consequently, efforts to reduce the computational burden have resulted in an extensive literature detailing many approaches, e.g. process convolution (Higdon, 2002), fixed rank kriging (Cressie and Johannesson, 2008), Gaussian predictive process models (Banerjee *et al.*, 2008), covariance tapering (Furrer *et al.*, 2006) and approximation by a GMRF (Rue and Tjelmeland, 2002).

Fitting an areal mixed model can also require expensive matrix operations. It is well known that a univariate Metropolis–Hastings algorithm for sampling from the posterior distribution of **W** leads to a slow mixing Markov chain because the components of **W** exhibit strong *a posteriori* dependence. This has led to various approaches that involve updating the random effects in a block(s). Constructing proposals for these updates is challenging, and the improved mixing comes at the cost of increased running time per iteration (see, for instance, Knorr-Held and Rue (2002), Haran *et al.* (2003) and Haran and Tierney (2010)).

The random effects for the RHZ model and for our sparse model, in contrast, are practically *a posteriori* uncorrelated. This means that we can use a spherical normal proposal for the random effects, which is very efficient computationally. Thus the computational crux of fitting the RHZ or sparse model is respectively the evaluation of the quadratic form $\boldsymbol{\delta}'\mathbf{Q}_R\boldsymbol{\delta}$ or $\boldsymbol{\delta}_S'\mathbf{Q}_S\boldsymbol{\delta}_S$. This oper-

ation has time complexity $O(n^2)$ for the RHZ model, which is sufficient to discourage or prevent the application of the model to large data sets. As we shall show in the next section, evaluation of $\boldsymbol{\delta}_S' \mathbf{Q}_S \boldsymbol{\delta}_S$ can be $O(1)$, which renders our sparse model applicable to even very large data sets.

6. Simulated application

We applied the classical generalized linear model and the three spatial models to binary, count and normal data simulated from the sparse model. The underlying graph for the binary and count data was the 30×30 lattice, and we restricted the co-ordinates of the vertices to the unit square. We chose for our design matrix $\mathbf{X} = [\mathbf{x} \quad \mathbf{y}]$, where $\mathbf{x} = (x_1, \ldots, x_{900})'$ and $\mathbf{y} = (y_1, \ldots, y_{900})'$ are the x- and y-co-ordinates of the vertices, and we let $\boldsymbol{\beta} = (1, 1)'$. A level plot of this large-scale structure, $\mathbf{X}\boldsymbol{\beta} = (x_1 + y_1, \ldots, x_{900} + y_{900})'$, is shown in Fig. 2.

Fig. 3 shows the standardized eigenvalues for the 30×30 lattice. Since over half of the values are non-positive, we chose to use only the first 400 eigenvectors to simulate data for our study, i.e. $\dim(\boldsymbol{\delta}_S) = 400$ and $\mathbf{M}$ is 900×400. The horizontal grey line marks the 400th eigenvalue, which equals 0.05.

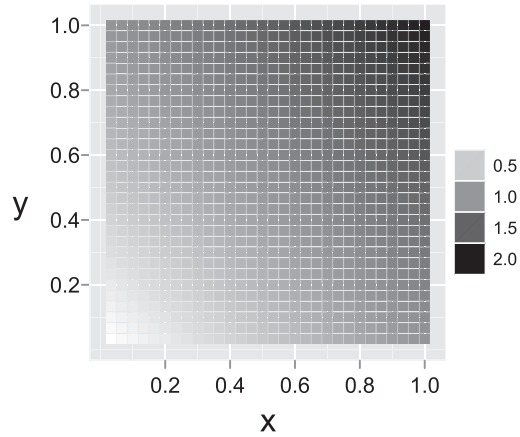

Fig. 2. Large-scale structure $\mathbf{X}\boldsymbol{\beta}$ for our simulation study

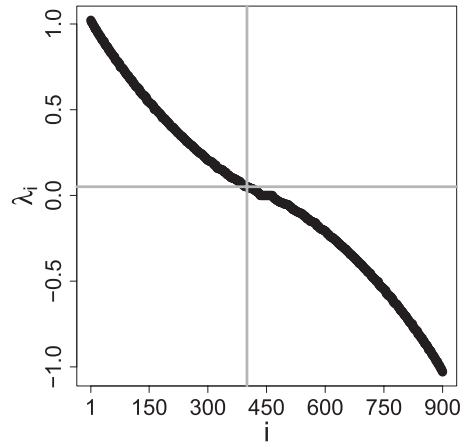

Fig. 3. Standardized eigenvalues for the 30 × 30 lattice

We applied all three spatial models in a Bayesian setting. It is customary to put a flat prior or diffuse spherical normal prior on $\boldsymbol{\beta}$. We assumed that $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 100\mathbf{I})$. The choice of a prior for τ has been the subject of debate. Kelsall and Wakefield (1999) recommended a gamma prior with shape parameter 0.5 and scale parameter 2000. This prior is appealing because it corresponds to the prior belief that the fixed effects are sufficient to explain the data (since a large value for τ implies small variances for the random effects) and because it does not produce artefactual spatial structure in the posterior.

To ensure that the Monte Carlo standard errors were sufficiently small, we simulated sample paths of length 2 million in all cases (Flegal *et al.*, 2008; Jones *et al.*, 2006). We fit the SGLMMs by using Metropolis–Hastings random-walk updates and/or Gibbs updates. For the normal data we used Gibbs updates for all parameters, and we used a Gibbs update for τ for all three types of data. For the binary and count data we updated $\boldsymbol{\beta}$ by using a random walk with proposal $\boldsymbol{\beta}^{(j+1)} \sim \mathcal{N}(\boldsymbol{\beta}^{(j)}, \hat{\mathbf{V}})$, where $\hat{\mathbf{V}}$ is the estimated asymptotic covariance matrix from a classical generalized linear model fit. We updated $\mathbf{W}$ by using a univariate random walk with normal proposals. And we updated each of $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_S$ by using a multivariate random walk with a spherical normal proposal.

6.1. Binary data

We created a binary data set by first setting $\tau = 1$ and simulating random effects according to $\boldsymbol{\delta}_S \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_S^{-1})$. Then we simulated independent observations according to $\mathbf{Z}|\boldsymbol{\delta}_S \sim \mathcal{B}(\mathbf{p})$, where $\mathcal{B}$ denotes a Bernoulli random variable and $\mathbf{p} = \exp(\mathbf{x} + \mathbf{y} + \mathbf{M}\boldsymbol{\delta}_S)/\{1 + \exp(\mathbf{x} + \mathbf{y} + \mathbf{M}\boldsymbol{\delta}_S)\}$ is the vector of true probabilities.

We fitted the simulated data set by using a non-spatial model, i.e. the standard logistic model, the centred autologistic model, the traditional SGLMM, the RHZ model and our sparse model with varying degrees of sparsity—400 eigenvectors (the true model), 200 eigenvectors, 100, 50 and 25. The study results are shown in Table 2, and Fig. 4 illustrates the inference for $\boldsymbol{\beta}$ with boxplots.

We see that the RHZ model and the true model produced approximately the same inference for these data. This is not surprising since the RHZ model can also fit residual structure and has random effects that essentially 'contain' the random effects for our sparse model, i.e. $\mathbf{L}\boldsymbol{\delta}$ can accommodate any structure that is exhibited by $\mathbf{M}\boldsymbol{\delta}_S$. The traditional SGLMM, in contrast, gave a rather poor point estimate of $\boldsymbol{\beta}$ along with confidence intervals that are over four times as wide as those provided by the other SGLMMs.

Table 2. Results of our analyses of a data set simulated from the sparse SGLMM with Bernoulli first stage†

Model	*Dimension*	$\hat{\beta}_1$	*CI*(β_1)	$\hat{\beta}_2$	*CI*(β_2)	$\hat{\tau}$	*CI*(τ)	$\|\mathbf{p} - \hat{\mathbf{p}}\|$	*Time (min)*
Non-spatial	—	0.822	(0.445, 1.199)	0.852	(0.475, 1.230)	—	—	4.576	—
Autologistic	—	0.930	(0.446, 1.397)	0.915	(0.448, 1.367)	—	—	3.876	20
Traditional	900	2.149	(0.254, 4.107)	2.218	(0.391, 4.132)	2.803	(0.287, 6.729)	3.255	2455
RHZ	898	1.079	(0.626, 1.538)	0.969	(0.536, 1.410)	0.949	(0.481, 1.482)	3.263	1640
Sparse	400	1.013	(0.569, 1.465)	0.946	(0.520, 1.382)	1.425	(0.355, 2.477)	3.174	285
Sparse	200	1.014	(0.582, 1.452)	0.937	(0.519, 1.359)	1.052	(0.365, 1.872)	3.175	167
Sparse	100	0.997	(0.566, 1.426)	0.923	(0.510, 1.342)	0.935	(0.339, 1.724)	3.201	80
Sparse	50	0.982	(0.562, 1.403)	0.900	(0.495, 1.304)	0.945	(0.317, 1.751)	3.295	49
Sparse	25	0.964	(0.552, 1.377)	0.897	(0.501, 1.301)	0.867	(0.225, 1.699)	3.543	37

†Here, and in Tables 4 and 5, 'traditional' refers to the traditional SGLMM.

Fig. 4. Boxplots illustrating inference for β for the simulated binary data: (a) β_1; (b) β_2

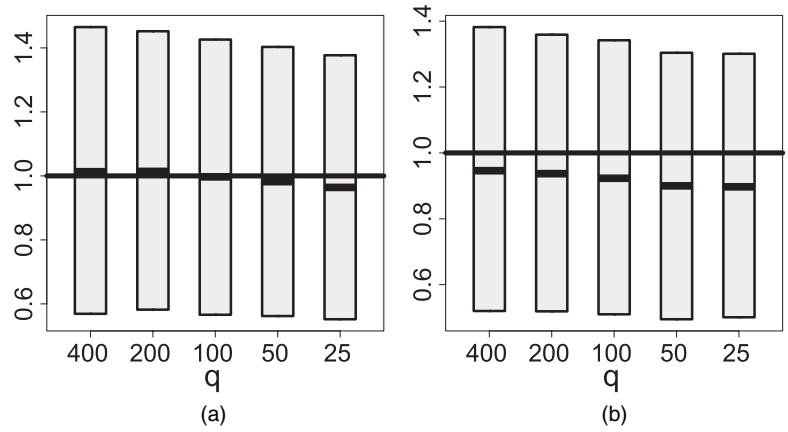

Fig. 5. Boxplots illustrating the effect of increasing sparsity on regression inference for the binary data: (a) β_1; (b) β_2

The boxplots in Fig. 5 and the level plots in Fig. 6 show the effect of increasing sparsity on regression inference and on fit respectively. The inference and fit that are provided by our model did not suffer appreciably until the number of eigenvectors had been reduced to 25, which represents a 97% reduction in the number of parameters relative to the traditional and RHZ models.

Fig. 7 shows the distributions of the estimated posterior correlations between the random effects for the three SGLMMs. We see that the random effects for the RHZ model and for our sparse model are at most weakly correlated, whereas the random effects for the traditional model are often strongly dependent. This is why spherical normal proposals are sufficient for $\boldsymbol{\delta}$ and $\boldsymbol{\delta}_S$.

To compare the performance of the three mixed models more throughly, we applied the models to 100 simulated binary data sets. The results are shown in Table 3. We see that the variance inflation caused by the traditional SGLMM was so large as to result in large type II error rates. Indeed, almost all of the 95% highest posterior density intervals covered 0. The RHZ and sparse models, in contrast, performed comparably and reliably.

Table 3. Results of applying the traditional, RHZ and sparse models to 100 simulated binary data sets†

Model	*Results for* β_1		*Results for* β_2	
	Coverage rate (%)	*Type II rate (%)*	*Coverage rate (%)*	*Type II rate (%)*
Traditional	94 ± 2	83 ± 4	95 ± 2	82 ± 4
RHZ	95 ± 2	0	97 ± 2	0
Sparse	95 ± 2	0	93 ± 3	0

†The sparse model was applied with $q = 50$.

6.2. Count data

We created a count data set by first simulating random effects according to $\boldsymbol{\delta}_S \sim \mathcal{N}\{\mathbf{0}, (3\mathbf{Q}_S)^{-1}\}$ and then simulating independent observations according to $\mathbf{Z}|\boldsymbol{\delta}_S \sim \mathcal{P}(\boldsymbol{\lambda})$, where $\mathcal{P}$ denotes a Poisson random variable and $\boldsymbol{\lambda} = \exp(\mathbf{x} + \mathbf{y} + \mathbf{M}\boldsymbol{\delta}_S)$ is the vector of true rates.

The study results for the count data are shown in Table 4. The sparse and RHZ models once again produced comparable inference. The traditional SGLMM produced a good point estimate of $\boldsymbol{\beta}$ for this data set but once again gave much wider intervals than the other models (Fig. 8).

6.3. Normal data

For the normal data we used the 20×20 lattice with the co-ordinates of the vertices again restricted to the unit square, and we used the first 180 eigenvectors of the corresponding Moran operator. We simulated random effects according to $\boldsymbol{\delta}_S \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_S^{-1})$ and then simulated observations according to $\mathbf{Z}|\boldsymbol{\delta}_S \sim \mathcal{N}(\boldsymbol{\mu} = \mathbf{x} + \mathbf{y} + \mathbf{M}\boldsymbol{\delta}_S, \sigma^2\mathbf{I})$, where $\sigma^2 = 1$.

The study results for these data are shown in Table 5. The RHZ and sparse models gave narrower intervals than the ordinary linear model because the latter model overestimated σ^2 (Fig. 9). For this data set the traditional SGLMM gave a poor estimate of $\boldsymbol{\beta}$ and inflated the variance so much as to cause a type II error. We note that the running times for this study were much longer than they would have been for binary and count data sets of the same size because we did a Gibbs update of the random effects for the normal data, which requires an expensive Cholesky decomposition.

6.4. On further dimension reduction

We conducted a more extensive simulation study with the aim of providing a heuristic for unsupervised dimension reduction, i.e. dimension reduction that does not employ the data to be analysed. For this study we used the 50×50 lattice, which is large but not so large as to make fitting the true model infeasible. We used the same regression component as before, i.e. $\mathbf{X} = [\mathbf{x}\,\mathbf{y}]$ and $\boldsymbol{\beta} = (1, 1)'$, and simulated binary data sets for four values of the smoothing parameter τ: 0.5, 1, 2 and 4. We used the first 1100 eigenvectors of the Moran operator, which correspond to standardized eigenvalues greater than 0.06 and thus practically exhaust the patterns of positive dependence for this scenario. We then fitted the true model and four sparser models to the simulated data. The number of random effects for the sparser models were 625 (the upper quartile of the eigenvalues), 265 (standardized eigenvalues greater than 0.7), 100 and 50. The last two of these were chosen with the hope of recommending a small fixed dimension for the random effects.

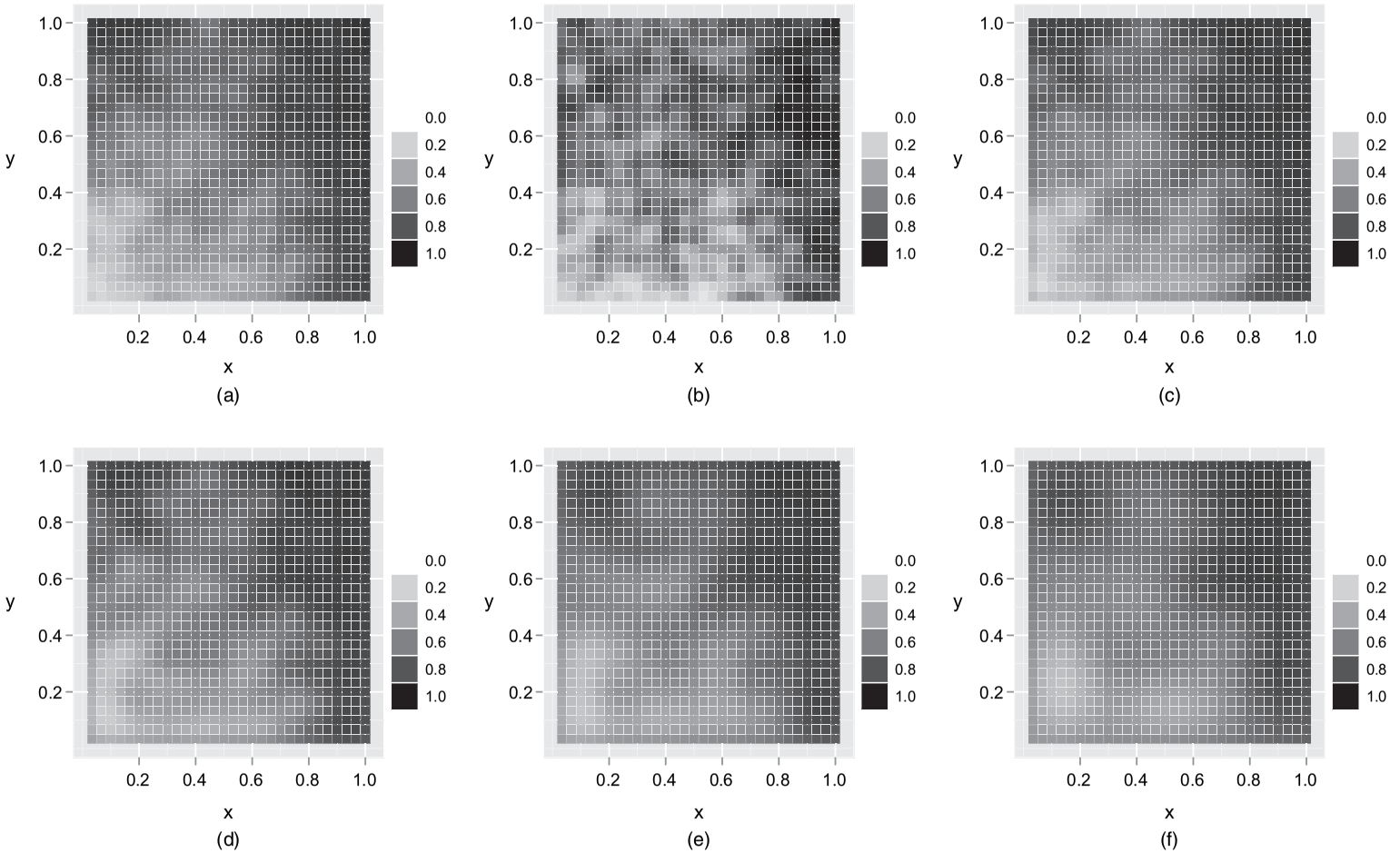

Fig. 6. True and fitted probabilities for the simulated binary data, for various values of $q = \dim(\delta_S)$: (a) fitted (q = 400); (b) true; (c) fitted (q = 200); (d) fitted (q = 100); (e) fitted (q = 50); (f) fitted (q = 25)

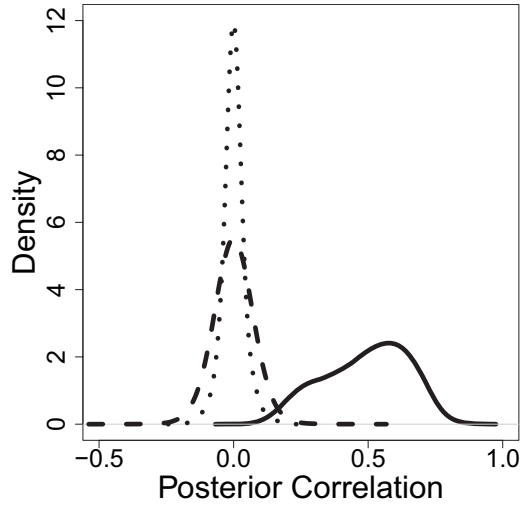

Fig. 7. Posterior correlations between the random effects for the three SGLMMs: ———, traditional; – – –, RHZ; · · · · · · · , sparse

Table 4. Results of our analyses of a data set simulated from the sparse SGLMM with Poisson first stage

Model	*Dimension*	$\hat{\beta}_1$	$CI(\beta_1)$	$\hat{\beta}_2$	$CI(\beta_2)$	$\hat{\tau}$	$CI(\tau)$	$\|\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}}\|$	*Time (min)*
Non-spatial	—	1.108	(0.989, 1.226)	1.179	(1.061, 1.297)	—	—	53.898	—
Traditional	900	1.007	(0.073, 1.913)	1.056	(0.117, 1.986)	5.141	(3.469, 6.952)	26.039	2275
RHZ	898	0.934	(0.816, 1.051)	1.092	(0.975, 1.210)	4.978	(2.916, 7.324)	26.172	1864
Sparse	400	0.942	(0.824, 1.060)	1.101	(0.986, 1.215)	4.499	(2.873, 6.402)	24.379	485

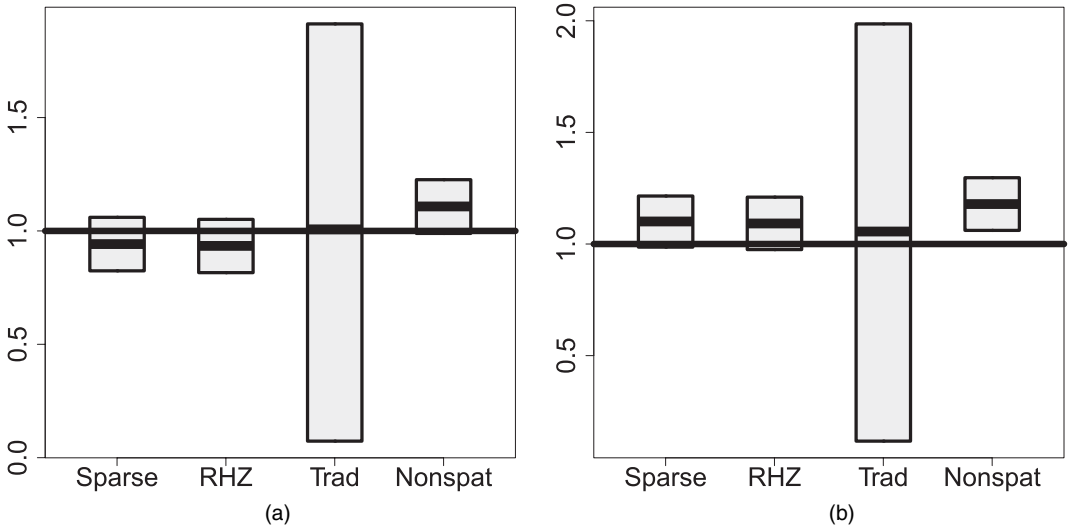

Fig. 8. Boxplots illustrating inference for β for the simulated count data: (a) β_1; (b) β_2

Table 5. Results of our analyses of a data set simulated from the sparse SGLMM with Gaussian first stage

Model	*Dimension*	$\hat{\beta}_1$	$CI(\beta_1)$	$\hat{\beta}_2$	$CI(\beta_2)$	$\hat{\tau}$	$CI(\tau)$	$\hat{\sigma}^2$	$CI(\sigma^2)$	$\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|$	*Time (min)*
Non-spatial	—	1.143	(0.824, 1.462)	0.925	(0.606, 1.244)	—	—	1.558	—	14.204	—
Traditional	400	1.583	(−0.816, 3.984)	0.288	(−2.066, 2.656)	0.825	(0.417, 1.322)	0.793	(0.585, 1.312)	9.798	1248
RHZ	398	1.143	(0.912, 1.377)	0.925	(0.692, 1.158)	0.847	(0.427, 1.365)	0.802	(0.595, 1.141)	9.725	1660
Sparse	180	1.143	(0.884, 1.403)	0.925	(0.666, 1.185)	0.961	(0.525, 1.486)	1.020	(0.866, 1.229)	8.692	385

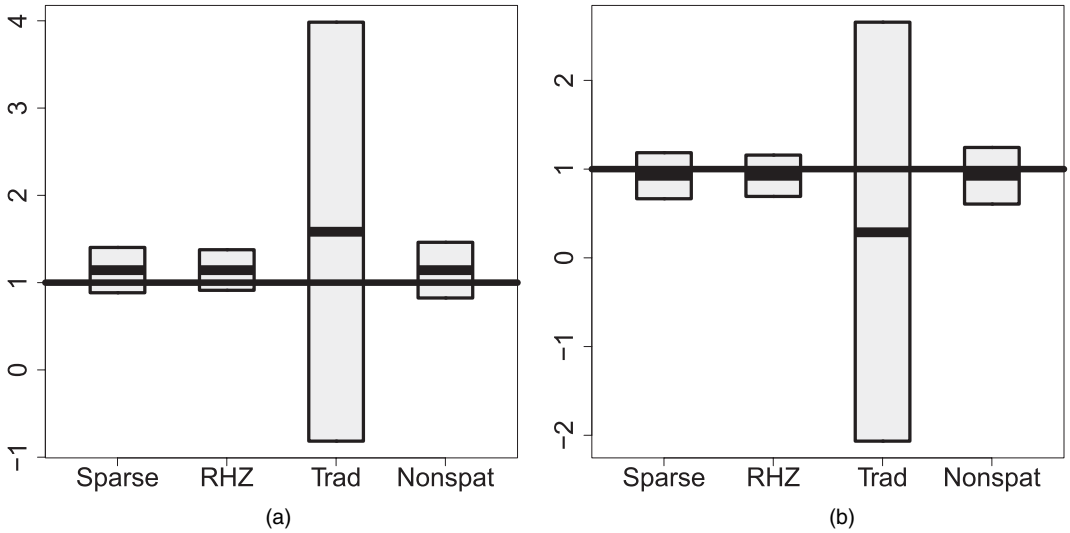

Fig. 9. Boxplots illustrating inference for β for the simulated Gaussian data: (a) β_1; (b) β_2

Table 6. Effect of various levels of sparsity, for simulated binary data with $\tau = 0.5, 1, 2, 4$

τ	*Dimension*	$\hat{\beta}_1$	$CI(\beta_1)$	$\hat{\beta}_2$	$CI(\beta_2)$	$\hat{\tau}$	$CI(\tau)$	$\|\mathbf{p}-\hat{\mathbf{p}}\|$	*Time*
0.5	1100	0.862	(0.581, 1.147)	1.166	(0.865, 1.463)	0.665	(0.439, 0.912)	6.379	5159
	625	0.869	(0.592, 1.147)	1.114	(0.828, 1.402)	0.617	(0.372, 0.915)	6.521	1836
	265	0.857	(0.588, 1.127)	1.085	(0.805, 1.367)	0.512	(0.311, 0.720)	6.877	597
	100	0.852	(0.588, 1.117)	1.035	(0.766, 1.307)	0.464	(0.253, 0.676)	7.424	260
	50	0.832	(0.571, 1.092)	1.023	(0.757, 1.291)	0.410	(0.206, 0.621)	7.853	113
1	1100	1.125	(0.870, 1.380)	0.992	(0.743, 1.251)	0.930	(0.500, 1.319)	5.205	6067
	625	1.097	(0.847, 1.344)	0.971	(0.724, 1.221)	1.056	(0.585, 1.687)	5.294	1666
	265	1.086	(0.840, 1.337)	0.962	(0.715, 1.212)	0.793	(0.404, 1.198)	5.418	579
	100	1.050	(0.809, 1.293)	0.924	(0.683, 1.165)	1.141	(0.533, 1.865)	5.787	169
	50	1.037	(0.796, 1.275)	0.920	(0.679, 1.161)	1.031	(0.435, 1.726)	6.005	109
2	1100	1.046	(0.800, 1.293)	0.988	(0.739, 1.239)	2.234	(0.676, 3.789)	4.134	4469
	625	1.046	(0.805, 1.291)	0.992	(0.743, 1.237)	1.701	(0.856, 2.715)	4.098	1926
	265	1.034	(0.793, 1.276)	0.990	(0.744, 1.238)	1.507	(0.548, 2.723)	4.116	552
	100	1.018	(0.780, 1.254)	0.966	(0.726, 1.210)	2.031	(0.616, 4.163)	4.230	184
	50	1.012	(0.779, 1.250)	0.960	(0.721, 1.202)	2.478	(0.465, 5.205)	4.392	131
4	1100	0.978	(0.743, 1.214)	1.173	(0.934, 1.411)	4.530	(1.192, 6.581)	3.238	4698
	625	0.986	(0.750, 1.224)	1.142	(0.903, 1.380)	3.470	(1.246, 7.501)	3.417	2316
	265	0.972	(0.733, 1.205)	1.133	(0.898, 1.374)	2.839	(1.103, 5.186)	3.506	546
	100	0.957	(0.725, 1.195)	1.126	(0.886, 1.363)	2.639	(0.935, 4.939)	3.567	249
	50	0.951	(0.716, 1.185)	1.122	(0.885, 1.360)	2.296	(0.697, 4.580)	3.652	155

The results of the study are shown in Table 6 and Fig. 10. Evidently, a very dramatic reduction in dimension comes at little cost to regression inference. In fact, since β_1 are β_2 are negatively correlated in this scenario, greater dimension reduction resulted in a better point estimate for $\boldsymbol{\beta}$ in many cases. And even using a mere 50 eigenvectors did not result in a large relocation of any highest posterior density interval. Hence, 50–100 eigenvectors may suffice for most analyses, which would remove evaluation of the above-mentioned quadratic form as the chief computational burden in fitting the model. Should one wish to be more conservative, we

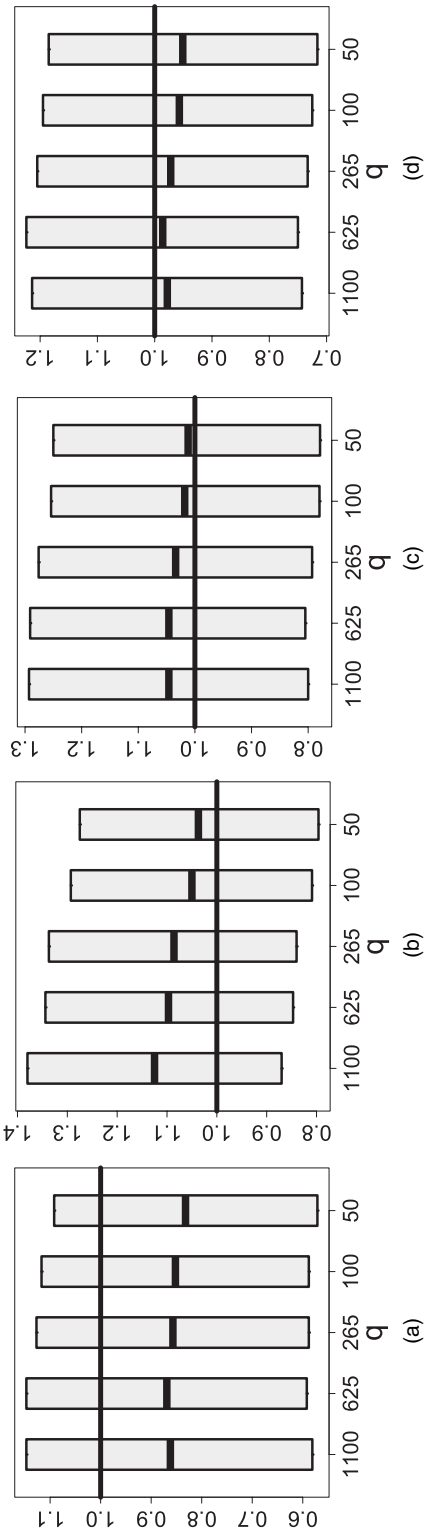

Fig. 10. Boxplots illustrating inference for β for simulated binary data with various values for τ and various degrees of sparsity: (a) β_1, $\tau = 0.5$; (b) β_1, $\tau = 1$; (c) β_1, $\tau = 2$; (d) β_1, $\tau = 4$; (e) β_2, $\tau = 0.5$; (f) β_2, $\tau = 1$; (g) β_2, $\tau = 2$; (h) β_2, $\tau = 4$

recommend the use of all eigenvectors corresponding to standardized eigenvalues that are greater than 0.7. For a square lattice this means using approximately 10% of the eigenvectors.

For a more principled approach to further dimension reduction, we recommend fitting a sequence of models and using the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002) to choose between them. For each of the four data sets just described, the model with $q = 50$ yielded the smallest value of the DIC. And, for the simulated binary data analysed in Section 6.1, the model with $q = 25$ yielded the smallest value for the DIC. These results lend further support to our claim that 50–100 eigenvectors may suffice for most analyses, but this DIC-based approach to selecting q is obviously more defensible and is only slightly more burdensome computationally.

7. Application to US infant mortality data

The plot in Fig. 11 shows infant mortality data for 3071 US counties. Each shaded circle represents a ratio of deaths to births, i.e. an empirical infant mortality rate, for a given county. The data were obtained from the 2008 area resource file, which is a county level database maintained by the Bureau of Health Professions, Health Resources and Services Administration, US Department of Health and Human Services. Specifically, three variables were extracted from the area resource file: the 3-year (2002–2004) average number of infant deaths before the first birthday, the 3-year average number of live births and the 3-year average number of low birth weight infants.

To these data we fit our sparse Poisson SGLMM with

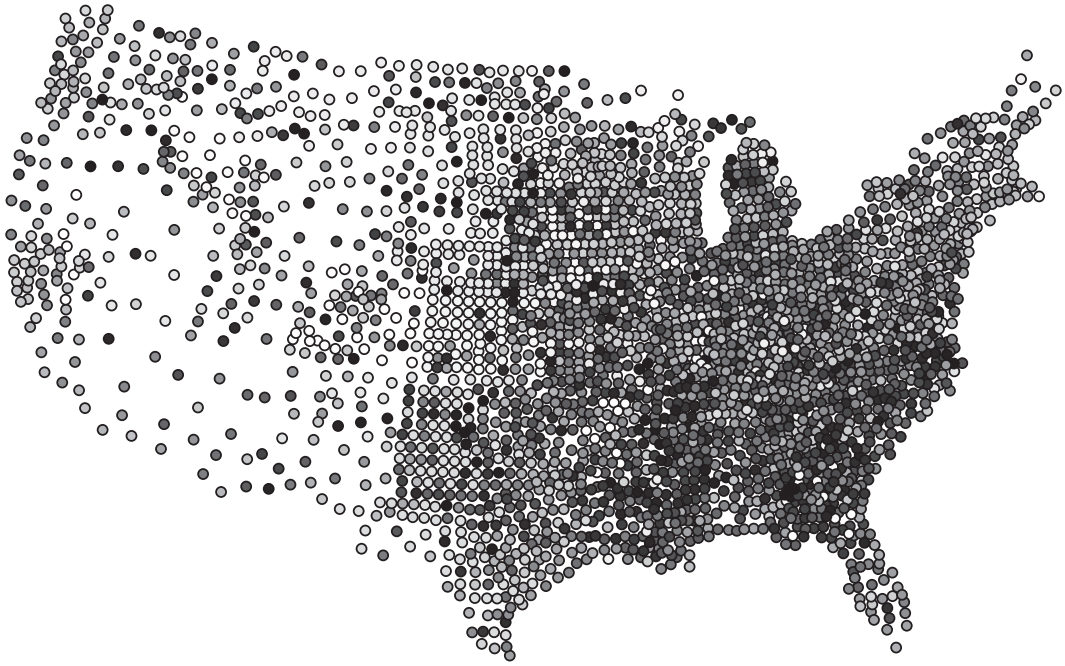

Fig. 11. Plot of infant mortality rates for 3071 counties of the contiguous USA: the circle for a given county represents that county's ratio of deaths to births, where births and deaths were averaged over the 3-year period 2002–2004; a darker circle indicates a higher rate

Table 7. Results of fitting model (4) to the infant mortality data

Predictor	*Parameter*	*Estimate*	*CI*
Intercept	β_0	−5.430	(−5.616, −5.246)
Low birth weight	β_1	8.777	(7.540, 10.032)
Black	β_2	0.00423	(0.00288, 0.00556)
Hisp	β_3	−0.00379	(−0.00488, −0.00270)
Gini	β_4	−0.555	(−0.977, −0.125)
aff	β_5	−0.0757	(−0.0877, −0.0638)
stab	β_6	−0.0285	(−0.0433, −0.0138)
—	τ	9.540	(3.870, 16.459)

$$\mathbb{E}(\text{deaths}_i|\boldsymbol{\beta}, \boldsymbol{\delta}_S) = \text{births}_i \exp(\beta_0 + \beta_1 \text{low}_i + \beta_2 \text{black}_i + \beta_3 \text{Hisp}_i + \beta_4 \text{Gini}_i + \beta_5 \text{aff}_i + \beta_6 \text{stab}_i + \mathbf{M}_i \boldsymbol{\delta}_S), \quad (4)$$

where deaths is the number of infant deaths, births is the number of live births, low is the rate of low birth weight, black is the percentage of black residents (according to the 2000 US census), Hisp is the percentage of Hispanic residents (2000 US census), Gini is the Gini coefficient, which is a measure of income inequality (Gini, 1921), aff is a composite score of social affluence (Yang *et al.*, 2009) and stab is residential stability, which is an average z-score of two variables from the 2000 US census.

We note that $\mathbf{M}$ in equation (4) contains eigenvectors not of $\mathbf{P}^{\perp}\mathbf{A}\mathbf{P}^{\perp}$ but of $\mathbf{P}_w^{\perp}\mathbf{A}\mathbf{P}_w^{\perp}$, where $\mathbf{P}_w^{\perp} = \mathbf{I} - \mathbf{W}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{1/2}$. The matrix $\mathbf{W}$ is diagonal with $\mathbf{W}_{ii}$ a weight for the ith observation (Reich *et al.*, 2006).

We chose to fit models with q equal to 25, 50, 100 and 200. Since we fit the four models simultaneously, the analysis took just 8.5 h, which was the time required to fit the model with $q = 200$. The model for $q = 50$ had the smallest value of the DIC. This represents a 98% reduction of the model dimension.

Fitting the RHZ model to these data took approximately 14 days, partly because storing the sample paths for the 3064 random effects required the use of a 46 Gbyte file-backed matrix (Kane and Emerson, 2010). If we could have stored all the results in random-access memory, fitting the RHZ model would have taken approximately 4 days.

Our sparse model and the RHZ model yielded nearly the same inference. The results for the sparse fit are given in Table 7. We see that all predictors were found to be significant. The posterior mean for τ is much smaller than the prior mean of 1000, which contraindicates a non-spatial analysis of these data. Moreover, five eigenvectors (all of which were among the first 40) were found to be significant predictors.

8. Discussion

In this paper we developed a framework for reparameterization of the areal SGLMM. By accounting appropriately for the covariates of interest, our approach alleviates the spatial confounding that plagues the traditional SGLMM. And, by accounting for the underlying graph, our model includes patterns of positive spatial dependence only, which allows for dramatic reduction in the dimension of the random effects.

It is conceivable that some data sets, although rare, may exhibit residual patterns of spatial repulsion, in which case prediction would suffer for such a data set if one applied our sparse

model with $\lambda_q > 0$. But this scenario is easy to avoid—simply select eigenvectors from both ends of the spectrum. If we desire $\dim(\boldsymbol{\delta}_S) = q$, we could choose the first $q/2$ eigenvectors and the last $q/2$ eigenvectors, for example. We could fit a sequence of models corresponding to a sequence of values for q and use DIC to choose the 'right' model. This symmetric scheme would still result in a considerable dimension reduction, but the reduction would clearly not be as dramatic as that provided by the scheme that employs only the positive end of the spectrum. It is easy, though, to envision a more sophisticated scheme that would achieve the optimal reduction (for a typical data set) by eventually discarding all the repulsive eigenvectors.

We studied the performance of our approach for three of the most commonly used first-stage models: Bernoulli, Poisson and Gaussian. In all three cases our approach resulted in better regression inference and in considerable gains in the computational efficiency of the Markov chain Monte Carlo algorithms used for inference. A follow-up study indicates that a small fixed number of random effects is sufficient for all data sets. More conservatively, our sparse model appears to require no more than $0.1n$ random effects to perform well with respect to both regression and prediction. The resulting gain in computational efficiency will permit the relatively rapid analysis of data sets that were once considered too large for the areal SGLMM.

Reich *et al.* (2006) noted that their approach is valid not only for the intrinsic conditionally auto-regressive model but also for the areal model that uses a proper conditional auto-regression and for the point level model that uses a Gaussian process. Consequently, our approach also carries over to these cases. But the precision matrix of the proper conditional auto-regression and the covariance matrix of the Gaussian process both involve additional parameters. Because these parameters enter the computations in a complicated way, restricting the random effects (in a naive way, at least) will improve regression inference but will not permit such a dramatic gain in computational efficiency. Extending our framework to handle these models in a computationally efficient way will be the focus of a future investigation.

Acknowledgements

We thank Tse-Chuan Yang and Jim Hodges for helpful discussions, and Tse-Chuan Yang for providing the infant mortality data set. And we are grateful for input from the Joint Editor, Associate Editor and two referees; their thoughtful comments led to a much improved paper.

References

Assunção, R. and Krainski, E. (2009) Neighbourhood dependence in Bayesian spatial models. *Biometr. J.*, **51**, 851–869.

Banerjee, S., Carlin, B. and Gelfand, A. (2004) *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: Chapman and Hall.

Banerjee, S., Gelfand, A. E., Finley, A. O. and Sang, H. (2008) Gaussian predictive process models for large spatial data sets. *J. R. Statist. Soc.* B, **70**, 825–848.

Besag, J. and Kooperberg, C. (1995) On conditional and intrinsic autoregression. *Biometrika*, **82**, 733–746.

Besag, J., York, J. and Mollié, A. (1991) Bayesian image restoration, with two applications in spatial statistics. *Ann. Inst. Statist. Math.*, **43**, 1–20.

Boots, B. and Tiefelsdorf, M. (2000) Global and local spatial autocorrelation in bounded regular tessellations. *J. Geog. Syst.*, **2**, 319.

Caragea, P. and Kaiser, M. (2009) Autologistic models with interpretable parameters. *J. Agric. Biol. Environ. Statist.*, **14**, 281–300.

Christensen, O. and Waagepetersen, R. (2002) Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, **58**, 280–286.

Clayton, D., Bernardinelli, L. and Montomoli, C. (1993) Spatial correlation in ecological models. *Int. J. Epidem.*, **22**, 1193–1202.

Cressie, N. A. (1993) *Statistics for Spatial Data*, 2nd edn. New York: Wiley.
Cressie, N. and Johannesson, G. (2008) Fixed rank kriging for very large spatial data sets. *J. R. Statist. Soc.* B, **70**, 209–226.
De Oliveira, V. (2000) Bayesian prediction of clipped gaussian random fields. *Computnl Statist. Data Anal.*, **34**, 299–314.
Diggle, P., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with discussion). *Appl. Statist.*, **47**, 299–350.
Durbin, J. and Watson, G. (1950) Testing for serial correlation in least squares regression: I. *Biometrika*, **37**, 409–428.
Flegal, J., Haran, M. and Jones, G. (2008) Markov chain Monte Carlo: can we trust the third significant figure? *Statist. Sci.*, **23**, 250–260.
Furrer, R., Genton, M. and Nychka, D. (2006) Covariance tapering for interpolation of large spatial datasets. *J. Computnl Graph. Statist.*, **15**, 502–523.
Geary, R. C. (1954) The contiguity ratio and statistical mapping. *Incorp. Statistn*, **5**, 115–145.
Gini, C. (1921) Measurement of inequality of incomes. *Econ. J.*, **31**, 124–126.
Griffith, D. A. (2003) *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding through Theory and Scientific Visualization*. New York: Springer.
Haran, M. (2012) Gaussian random field models for spatial data. In *Markov Chain Monte Carlo Handbook* (eds S. P. Brooks, A. Gelman, G. L. Jones and X. L. Meng). Boca Raton: CRC Press. To be published.
Haran, M., Hodges, J. and Carlin, B. (2003) Accelerating computation in markov random field models for spatial data via structured mcmc. *J. Computnl Graph. Statist.*, **12**, 249–264.
Haran, M. and Tierney, L. (2010) On automating markov chain monte carlo for a class of spatial models. *Baysn Anal.*, to be published.
Higdon, D. (2002) Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues* (eds C. Anderson, V. Barnett, P. Chatwin and A. El-Shaarawi), pp. 37–56. London: Springer.
Hughes, J., Haran, M. and Caragea, P. (2011) Autologistic models for binary data on a lattice. *Environmetrics*, **22**, 857–871.
Jones, G., Haran, M., Caffo, B. and Neath, R. (2006) Fixed-width output analysis for markov chain Monte Carlo. *J. Am. Statist. Ass.*, **101**, 1537–1547.
Kaiser, M. and Cressie, N. (1997) Modeling poisson variables with positive spatial dependence. *Statist. Probab. Lett.*, **35**, 423–432.
Kaiser, M. and Cressie, N. (2000) The construction of multivariate distributions from markov random fields. *J. Multiv. Anal.*, **73**, 199–220.
Kane, M. J. and Emerson, J. W. (2010) bigmemory: manage massive matrices with shared memory and memory-mapped files. *R Package Version 4.2.3*.
Katznelson, Y. (2004) *An Introduction to Harmonic Analysis*. Cambridge: Cambridge University Press.
Kelsall, J. and Wakefield, J. (1999) Discussion of Bayesian models for spatially correlated disease and exposure data, by Best et al. In *Bayesian Statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith). Oxford: Oxford University Press.
Knorr-Held, L. and Rue, H. (2002) On block updating in markov random field models for disease mapping. *Scand. J. Statist.*, **29**, 597–614.
Lunn, D., Thomas, A., Best, N. and Spiegelhalter, D. (2000) Winbugs—a bayesian modelling framework: concepts, structure, and extensibility. *Statist. Comput.*, **10**, 325–337.
Moran, P. (1950) Notes on continuous stochastic phenomena. *Biometrika*, **37**, 17–23.
Nelder, J. A. and Wedderburn, R. W. M. (1972) Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.
Reich, B. J., Hodges, J. S. and Zadnik, V. (2006) Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, **62**, 1197–1206.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields: Theory and Applications*. Boca Raton: Chapman and Hall.
Rue, H. and Tjelmeland, H. (2002) Fitting gaussian markov random fields to gaussian fields. *Scand. J. Statist.*, **29**, 31–49.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Wall, M. (2004) A close look at the spatial structure implied by the car and sar models. *J. Statist. Planng Inf.*, **121**, 311–324.
Yang, T.-C., Teng, H.-W. and Haran, M. (2009) The impacts of social capital on infant mortality in the U.S.: a spatial investigation. *Appl. Spatl Anal. Poly*, **2**, 211–227.