ELSEVIER Computational Statistics & Data Analysis 26 (1997) 1–15

COMPUTATIONAL STATISTICS & DATA ANALYSIS

Sparse matrix tools for Gaussian models on lattices

S.P. Smith

EA Engineering, Science and Technology, 3468 Mt. Diablo Blvd., Suite B-100, Lafayette, CA 94549, USA

Received 1 April 1996; received in revised form 1 March 1997

Abstract

Sparse matrix methods are considered for Gaussian models on lattices. The approach depends on the applicability of mixed model methodology, which is feasible when observational errors are present. Moreover, lattice models with nearest-neighbor interactions lead to a spatial variance–covariance matrix having a sparse inverse, and this lends itself to sparse-matrix manipulation. Modern technology involving the Cholesky decomposition and backward differentiation are described for prediction, parameter estimation, cross validation, and conditional simulation of spatial processes. © 1997 Elsevier Science B.V.

Keywords: Backward differentiation; Conditional simulation; Cholesky decomposition; Mixed linear model; Nearest neighbor analysis; Nonparametric regression; Spatial covariance; Random fields; Restricted maximum likelihood; Variance components

1. Introduction

A lattice model is different from other spatial models, in that dependences involving adjacent elements around a grid point are built directly into the model, and reference to a covariance function is not made. A spatial variable on a rectangular lattice is denoted by u_{ij} corresponding to the ith row and jth column. Whittle (1954) presented the simultaneous bilateral model of the form

$$u_{ij} = \sum_{rs\varepsilon\Omega} \alpha_{rs} u_{i-rj-s} + \sigma\varepsilon_{ij}, \tag{1}$$

PII S0167-9473(97)00023-6

where the collection $\{\varepsilon_{ij}$: all i and $j\}$ are independent normal 0–1 deviates and Ω is a finite index set for r and s. In matrix form, this equation is presented as

$$\boldsymbol{u} = \Delta\boldsymbol{u} + \varepsilon \tag{2}$$

where $E\{\varepsilon\} =$ null and $\text{var}\{\varepsilon\} = \sigma^2\boldsymbol{I}$.

Variables on the edge of a finite lattice are undefined, because the summation in (1) includes neighboring cells outside the lattice. An ideal treatment for edges is unavailable, but some guidance is provided in Section 2.2. Nevertheless, the form of (2) implies that $\text{var}\{\boldsymbol{u}\}$ is $\sigma^2(\boldsymbol{I} - \Delta)^{-1}(\boldsymbol{I} - \Delta)^{-1'}$, and the inverse variance matrix is more simply $\sigma^{-2}(\boldsymbol{I} - \Delta)'(\boldsymbol{I} - \Delta)$ (Cressie, 1991, Section 6.7).

A different approach, the conditional formulation, was introduced by Besag (1974), in which (1) is given as the regression of u_{ij} on all other u_{i-rj-s} ($r \neq 0$ and $s \neq 0$). In this situation, the members of ε are no longer uncorrelated. Because Δ depicts partial regression coefficients, $\text{var}\{\boldsymbol{u}\}$ is necessarily $(\boldsymbol{I} - \Delta)^{-1}\boldsymbol{D}$ where $\boldsymbol{D}$ is a diagonal matrix containing the prediction error variances of the regressions (Cressie, 1991, Section 6.6).

Other techniques for lattices have been borrowed from time series analysis. Gleeson and Cullis (1987) present an ARMA model that is suitable for one-dimensional lattices. Cullis and Gleeson (1991) extended this approach to two dimensions by depicting the spatial error as a separable lattice process that combines the ARMA specifications for rows and columns. Basu and Reinsel (1993) described a first-order unilateral ARMA model, noting that a two-dimensional grid can be naturally represented as a partially ordered set in one dimension. The correct treatment of edge effects is less problematic when constructing the likelihood for these models, because statistical variation is propagated from initial conditions as characterized by time series.

The omission of an observational error or nugget effect is typical of some lattice models, but not all (e.g., Zimmerman, 1989a). Nugget effects have been introduced by geostatisticians in a popular statistical analysis called kriging, which is a process for smoothing spatial data by filtering out the random variation that is not correlated spatially (Laslett, 1994). To this end, mixed model methodology (cf. Harville, 1976) offers some advantages when there are prominent observational errors. The existence of an observational error permits the use of simplifying expressions involving the inverse of a matrix sum, and this leads to the mixed-model equations decribed in Section 3.1. Advances in sparse-matrix methods, particularly as applied to the Cholesky decomposition and its derivatives, have made the large number of grid cells typical of lattice models less of a burden on computational resources.

The purpose of this paper is to present novel applications of mixed model methodology for lattice models. Some lattice models are described in Section 2, and Section 3 provides a cursory treatment of the mixed linear model. Section 4 presents the associated sparse-matrix tools, including variance–covariance estimation and conditional simulation. An illustration of a sparse-matrix implementation for cross validation is given in Section 5.

2. Lattice structure

2.1. Neighborhood relationships

Lattice models are usually specified over a one or two-dimensional grid, but higher dimensions are allowable. In two dimensions, the grid may be rectangular (Fig. 1(a)), or it can be a network of hexagons (Fig. 1(b)), as used by Mead (1967). The measured objects are not the rectangles or hexagons, but the vertices, given by dark dots, and the nearest neighbor relationships are indicated by the solid connections. Each of the vertices in Fig. 1(b) are connected to three neighboring vertices (first order), and these are connected, in turn, to six vertices (second order). The nearest-neighbor relationships are given as in Eq. (2) for simultaneous models, but it is useful to separate the border effects as indicated below.

$$\boldsymbol{u}_1 = \Delta_{11}\boldsymbol{u}_1 + \Delta_{21}\boldsymbol{u}_2 + \varepsilon_1, \tag{3}$$

$$\boldsymbol{u}_2 = \Delta_{12}\boldsymbol{u}_1 + \Delta_{22}\boldsymbol{u}_2 + \varepsilon_2, \tag{4}$$

where $\boldsymbol{u}_1$ contains the effects on the border, $\boldsymbol{u}_2$ contains the effects interior to the border, and $\Delta_{ij}\ \{i,j = 1,2\}$ are matrices that depict the nearest-neighbor relationships selected from Δ. It is feasible to specify the lattice model using only the simultaneous relationships of (4), but this assumes the model is interpreted conditionally on $\boldsymbol{u}_1$. The mean vectors are taken as null, and the variance–covariance structure is assumed from $\mathrm{var}\{\boldsymbol{u}_1\} = \boldsymbol{F}$, $\mathrm{cov}\{\boldsymbol{u}_1\varepsilon_2'\} = \text{null}$, and $\mathrm{var}\{\varepsilon_2\} = \sigma^2\boldsymbol{I}$. Eq. (2) can now be recast by partitioning $\boldsymbol{u}$, Δ and ε as

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_1 \\ \boldsymbol{u}_2 \end{bmatrix}, \qquad \Delta = \begin{bmatrix} 0 & 0 \\ \Delta_{12} & \Delta_{22} \end{bmatrix}, \qquad \varepsilon = \begin{bmatrix} \boldsymbol{u}_1 \\ \varepsilon_2 \end{bmatrix}.$$

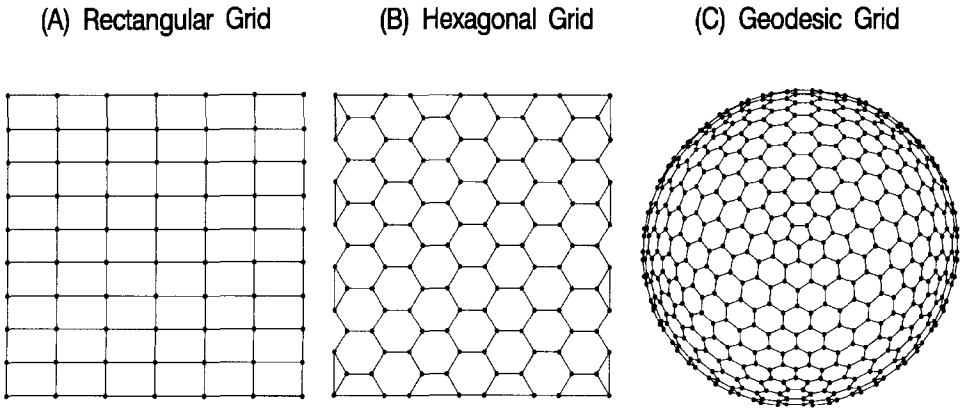

Fig. 1. Examples of element networks forming: (A) rectangular grid; (B) hexagonal grid; and (C) geodesic grid. Nearest neighbor relationships indicated by solid connections.

Therefore, (2) implies that the variance of $\boldsymbol{u}$ is $(\boldsymbol{I}-\Delta)^{-1}\mathrm{var}\{\varepsilon\}(\boldsymbol{I}-\Delta)^{-1'}$, and the inverse structure is more neatly:

$$\boldsymbol{G}^{-1}=\begin{bmatrix}\boldsymbol{F}^{-1}+\sigma^{-2}\Delta_{12}'\Delta_{12} & \sigma^{-2}\Delta_{12}'(\boldsymbol{I}-\Delta_{22}) \\ \sigma^{-2}(\boldsymbol{I}-\Delta_{22})'\Delta_{12} & \sigma^{-2}(\boldsymbol{I}-\Delta_{22})'(\boldsymbol{I}-\Delta_{22})\end{bmatrix}. \tag{5}$$

Aside from $\boldsymbol{F}^{-1}$, the matrix on the right of (5) is very sparse and is represented as a simple summation of rank-1 contributions from interior elements of the lattice.

It is possible to use both (3) and (4) to describe a simultaneous model in which the second moments are given as $\mathrm{var}\{\varepsilon_1\}=\mathscr{F}$, $\mathrm{var}\{\varepsilon_2\}=\sigma^2\boldsymbol{I}$, and $\mathrm{cov}\{\varepsilon_1\varepsilon_2'\}=$ null. For this situation the partitioning argument gives the inverse structure as:

$$\boldsymbol{G}^{-1}=\begin{bmatrix}\sigma^{-2}\Delta_{12}'\Delta_{12} & \sigma^{-2}\Delta_{12}'(\boldsymbol{I}-\Delta_{22}) \\ \sigma^{-2}(\boldsymbol{I}-\Delta_{22})'\Delta_{12} & \sigma^{-2}(\boldsymbol{I}-\Delta_{22})'(\boldsymbol{I}-\Delta_{22})\end{bmatrix}$$
$$+\begin{bmatrix}(\boldsymbol{I}-\Delta_{11})'\mathscr{F}^{-1}(\boldsymbol{I}-\Delta_{11}) & (\boldsymbol{I}-\Delta_{11})'\mathscr{F}^{-1}\Delta_{21} \\ \Delta_{21}'\mathscr{F}^{-1}(\boldsymbol{I}-\Delta_{11}) & \Delta_{21}'\mathscr{F}^{-1}\Delta_{21}\end{bmatrix}. \tag{6}$$

The second term of (6) only depicts a nonzero structure among elements in the few tiers that form the exterior of the lattice.

2.2. *Treatment of edge effects*

Edge effects can be treated by nominating $\boldsymbol{F}^{-1}$ or $\mathscr{F}^{-1}$, and using (5) or (6) in the matrix constructs defined in Section 3. It is feasible to take $\boldsymbol{F}^{-1}=\sigma^{-2}\boldsymbol{I}$ or $\mathscr{F}^{-1}=\sigma^{-2}\boldsymbol{I}$, effectively ignoring all or some of the neighborhood relationships along the border. This will induce some heteroscedasticity, but the approximation can be improved by extending the lattice beyond the borders. A different approach, using (5), is to represent the border elements as a closed loop (for first-order interactions), and model them separately as a spatial process without borders.

To treat border effects as fixed, as Gleeson and McGilchrist (1980) have done, simply set $\boldsymbol{F}^{-1}$ or $\mathscr{F}^{-1}$ to null. However, this should be done with caution, because (5) and (6) become singular, and accommodations are needed for the numerical protocol. Moreover, there must be enough information available near the border of the lattice that the loss in degrees of freedom is not extreme. Higher dimensional lattices are most sensitive to information loss along the border. The form of (5) and (6) may naturally be singular for certain nonstationary processes (see Section 5), but this does not necessarily imply a major loss of information due to a border.

The treatment of border effects is only difficult when there are edges. When there are no edges, the inverse structure is given by $\sigma^{-2}(\boldsymbol{I}-\Delta)'(\boldsymbol{I}-\Delta)$. Nevertheless, there is likely to be some fudging of the geometry. If small distortions on the border are acceptable, it is possible to connect border elements and effectively remove the border. An example is found in Fig. 1(b), where all elements are connected to three neighboring elements, even on the border. Some lattice structures naturally lack edges, and Cressie (1991, p. 438) suggested wrapping rectangular lattices in a torus. Cressie admits that this fix is arbitrary for rectangles. A lattice representing the

surface of a sphere, such as the earth, has no edge. Elements of a spherical lattice can be defined by a collection of hexagons and pentagons that constitute Buckminster Fuller's geodesic dome (Fig. 1(c)). The braces on the dome surface are of equal length, and they connect each vertex to three neighboring vertices, and this provides a first-order relationship.

3. Statistical theory

3.1. The mixed linear model

The mixed linear model is represented by

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is the vector of observations, $\boldsymbol{\beta}$ the vector of fixed effects, $\boldsymbol{u}$ the vector of random effects, and $\boldsymbol{e}$ the observational error.

The matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are incidence matrices that relate the various effects to observations. The first moments for the random effects are $E\{\boldsymbol{u}\} = \text{null}$ and $E\{\boldsymbol{e}\} = \text{null}$, and the variance–covariance structure is given by $\text{var}\{\boldsymbol{u}\} = \boldsymbol{G}$, $\text{var}\{\boldsymbol{e}\} = \boldsymbol{R}$, and $\text{cov}\{\boldsymbol{u}, \boldsymbol{e}'\} = \text{null}$. Additional assumptions are needed to implement maximum likelihood or computer simulation, and generally $\boldsymbol{y}$, $\boldsymbol{u}$, and $\boldsymbol{e}$ are taken as multivariate normal.

In the lattice model context, $\boldsymbol{u}$ are unobserved spatial effects that depict the dependence of adjacent elements in the lattice, and $\boldsymbol{G}$ is a variance matrix described in Section 2.1. The observation vector $\boldsymbol{y}$ reflects spatial dependence only indirectly through $\boldsymbol{u}$, and therefore the mixed model automatically treats missing values. The vector $\boldsymbol{e}$ is made up of random nugget effects, and $\boldsymbol{R} = \sigma_e^2 \boldsymbol{I}$. In spatial models, the vector $\boldsymbol{\beta}$ may contain treatment effects representing such things as fertilizers or crop varieties and also, possibly, fixed border effects.

In addition, there are useful ways to depict time series models that are applicable to one and two dimensional lattices. For example, state-space theory permits the representation of an ARMA(p, q) process by the linear combination, $\boldsymbol{Zu}$; simply take $\boldsymbol{Z}$ to contain the coefficients of a MA(q) process as indicated by Eq. (13.1.25) of Hamilton (1994), and let the elements of $\boldsymbol{u}$ depict an AR(p) process as indicated by Eq. (13.1.24). The matrix, $\boldsymbol{G}^{-1}$, is sparse and can be constructed directly when $\boldsymbol{u}$ is an AR(p) (Robinson, 1991). Moreover, $\boldsymbol{Zu}$ may represent a separable row and column process on a rectangular lattice. The variance matrix for $\boldsymbol{Zu}$ is the Kronecker product of the variance matrices of the one-dimensional processes (Martin, 1979). Therefore, for the ARMA row and column process, we have $\boldsymbol{Z} = \boldsymbol{Z}_c \otimes \boldsymbol{Z}_r$ and $\boldsymbol{G}^{-1} = \boldsymbol{G}_c^{-1} \otimes \boldsymbol{G}_r^{-1}$ were the subscripts indicate row and column specifications.

As indicated by Goldberger (1962), the best linear unbiased prediction (BLUP) of $\boldsymbol{u}$ is found by evaluating

$$\hat{\boldsymbol{u}} = \boldsymbol{GZ}'\boldsymbol{V}^{-1}[\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}],$$

where $\boldsymbol{V} = \text{var}(\boldsymbol{y}) = \boldsymbol{ZGZ}' + \boldsymbol{R}$, and $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimate (BLUE) of the fixed effects obtained by generalized least squares, or by solving

$$[\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{X}][\hat{\boldsymbol{\beta}}] = [\boldsymbol{X}'\boldsymbol{V}^{-1}\boldsymbol{y}]. \tag{7}$$

These equations can be reformulated so that the solutions can be obtained directly from the mixed model equations (Henderson et al., 1959):

$$\begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix} \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}.$$

To simplify the notation let

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}, \qquad \boldsymbol{r} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix},$$

$$\hat{\boldsymbol{b}} = \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\boldsymbol{u}} \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix}. \tag{8}$$

Therefore the mixed model equations can be written as:

$$\boldsymbol{C}\hat{\boldsymbol{b}} = \boldsymbol{r}.$$

It is well known that if a noninformative prior is used to describe the fixed effects in a Bayesian context, the posterior distribution (conditional on $\boldsymbol{y}$) of a linear combination of $\boldsymbol{b}$, say $\boldsymbol{Hb}$, is multivariate normal, having a mean vector $\boldsymbol{H}\hat{\boldsymbol{b}}$, and a variance–covariance matrix $\boldsymbol{HC}^{-1}\boldsymbol{H}'$ (e.g., Dempfle, 1977).

The mixed model equations are only useful when $\boldsymbol{R}^{-1}$ and $\boldsymbol{G}^{-1}$ can be found. A significant drawback to regular use of the mixed-model equations is the lack of direct methods for calculating the inverse of variance–covariance matrices that are typical to spatial fields, i.e., apart from the lattice structures implied by Eqs. (5) and (6). Vecchia (1992) developed an approximate orthogonal basis using nearest-neighbor analysis, and this leads to a sparse inverse. Zimmerman (1989b) provides an alternate algorithm for the inversion of block Toeplitz matrices, but this is less of a sparse-matrix tool and more of a tool for treating special structure.

When observation error is absent, as it is with some spatial and time series models, the mixed model equations are undefined and alternative approaches (such as the Kalman filter) provide more logical avenues of analysis. However, there are sparse-matrix tools that are suitable for the case when $\boldsymbol{R}$ is singular or null, and when $\boldsymbol{R}$ is nonsingular these tools show the mixed model equations as by-products of elementary row operations (Smith, 1997).

3.2. *Likelihood functions*

The log-likelihood for the multivariate normal is given by

$$\text{const.} - \tfrac{1}{2}\log|\boldsymbol{V}| - \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\beta)'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\beta). \tag{9}$$

The maximum likelihood estimates of β and the dispersion parameters (that define $\boldsymbol{R}$ and $\boldsymbol{G}$) are found by maximizing (9). Estimates of the dispersion parameters can

be badly biased by small-sample errors induced by the estimation of β. This is a serious problem when the dimension of β is large relative to the information available to estimate β. To overcome this problem, Patterson and Thompson (1971) introduced restricted maximum likelihood (REML), where dispersion parameters are found by maximizing

$$\text{const.} - \tfrac{1}{2}\log|\boldsymbol{V}| - \tfrac{1}{2}|\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X}| - \tfrac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'\boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{10}$$

where $\hat{\boldsymbol{\beta}}$ is found from Eq. (7). Likelihoods (9) and (10) can be used directly for some lattice models (Gleeson and Cullis, 1987; Zimmerman, 1989a; Basu and Reinsel, 1994) and for more general spatial models (Jones and Vecchia, 1993; Zimmerman and Harville, 1991). However, another form of the likelihood is sometimes more suitable for purposes of numerical manipulation. To show this likelihood, follow the Bayesian approach and treat β as random with mean vector $\boldsymbol{\mu}_\beta$ and variance matrix $\alpha\boldsymbol{I}$. This prior becomes noninformative as $\alpha \to \infty$. Therefore, the mean and variance of $\boldsymbol{y}$ are given as

$$E\{\boldsymbol{y}\} = \boldsymbol{X}\boldsymbol{\mu}_\beta, \qquad \operatorname{var}\{\boldsymbol{y}\} = \boldsymbol{V}_\alpha = \boldsymbol{W}\boldsymbol{Q}_\alpha\boldsymbol{W}' + \boldsymbol{R},$$

where

$$\boldsymbol{W} = [\boldsymbol{X} \quad \boldsymbol{Z}], \qquad \boldsymbol{Q}_\alpha = \begin{bmatrix} \alpha\boldsymbol{I} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G} \end{bmatrix}.$$

These can be plugged back into (9), while using expressions for the determinant and inverse of a matrix sum (Henderson and Searle, 1981),

$$|\boldsymbol{V}_\alpha| = |\alpha\boldsymbol{I}| \cdot |\boldsymbol{G}| \cdot |\boldsymbol{R}| \cdot |\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{Q}_\alpha^{-1}|,$$

$$\boldsymbol{V}_\alpha^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{W}(\boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{Q}_\alpha^{-1})^{-1}\boldsymbol{W}'\boldsymbol{R}^{-1}.$$

Upon letting $\alpha \to \infty$, the relevant part of the log-likelihood becomes

$$\text{const.} - \tfrac{1}{2}\log|\boldsymbol{R}| - \tfrac{1}{2}\log|\boldsymbol{G}| - \tfrac{1}{2}\log|\boldsymbol{C}| - \tfrac{1}{2}\boldsymbol{y}'\boldsymbol{V}_\infty^{-1}\boldsymbol{y}, \tag{11}$$

where $\boldsymbol{C} = \boldsymbol{W}'\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{Q}_\infty^{-1}$ is the coefficient matrix of the mixed model equation. Because $\boldsymbol{V}_\infty^{-1}\boldsymbol{X} = \boldsymbol{0}$, (11) is invariant to $\boldsymbol{\mu}_\beta$. A version of (11) that is suitable for singular $\boldsymbol{G}$ is found in Harville (1977).

In general, $|\boldsymbol{C}| = 0$ and (11) will not be defined. It is standard to remove singularities from $\boldsymbol{X}$, so that $|\boldsymbol{C}| > 0$, but this is primarily a theoretical device. Singularities can be treated automatically in the numerical protocol.

The utility of (11) is apparent through its connection with the *mixed model matrix*:

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{Z} & \boldsymbol{y}'\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}.$$

Graser, Smith and Tier (1987) were among the first to appreciate that application of standard matrix operations on $\boldsymbol{M}$, like the Cholesky decomposition, lead to the

evaluation of (11). Like the mixed model equations, construction of $\boldsymbol{M}$ assumes that $\boldsymbol{G}^{-1}$ and $\boldsymbol{R}^{-1}$ are available.

4. Uses of the Cholesky decomposition

4.1. Background and common uses

Matrices like $\boldsymbol{C}$ and $\boldsymbol{M}$ can be very large. Fortunately, they are also usually very sparse, and this permits application of methods designed for sparse matrices (George and Liu, 1981). Of particular importance is the Cholesky decomposition, and the aspects related to the computation of the Cholesky decomposition have recently been advanced by Ng and Peyton (1993a, b) and Rothberg and Gupta (1994). This technology is finding applications in several important areas, including linear and quadratic programming, circuit simulation, and structural analysis. The numerical complexity of Cholesky's method to solve structural equations is well studied, and is very similar to what can be expected for the mixed model equations with lattices models.

The Cholesky decomposition is emphasized merely because it has been proven practical. Other techniques, and associated matrix structures, deserve consideration. For example, the QR algorithms, as applied to a rectangular array, is also useful for fitting data to the mixed linear model and this has been demonstrated in the area of quantitative genetics (Hudson, 1986). Moreover, sparse-matrix tools for treating indefinite systems are also useful because of the connection between the mixed linear model and symmetric and indefinite matrices (Smith, 1997).

The common use of the Cholesky decomposition is in solving linear equations, and details can be found in standard text books (e.g., Golub and Van Loan, 1983). In the present context, solutions to the mixed model equation are obtained with the following instructions, and these can be fully implemented in sparse matrix mode:

(a) Evaluate $\boldsymbol{L}$, $\boldsymbol{LL}' = \boldsymbol{C}$.
(b) Solve $\boldsymbol{g}$ by forward substitution, $\boldsymbol{Lg} = \boldsymbol{r}$,
(c) Solve $\hat{\boldsymbol{b}}$ by backward substitution, $\boldsymbol{L}'\hat{\boldsymbol{b}} = \boldsymbol{g}$.

To treat singularities, set the ith element of $\boldsymbol{g}$ and $\hat{\boldsymbol{b}}$ to zero whenever the ith diagonal of $\boldsymbol{L}$ is zero. Matrix $\boldsymbol{C}$ is required to be a non-negative definite matrix, which is the case when (8) applies.

4.2. Restricted maximum likelihood

Section 3.2 hints that the Cholesky decomposition is useful for REML applications. Two of the terms in the likelihood (11) are a function of the Cholesky decomposition of the mixed model matrix; i.e., $f(\boldsymbol{L}) = -\frac{1}{2}\boldsymbol{y}'\boldsymbol{V}_{\infty}^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{C}|$, where $\boldsymbol{LL}' = \boldsymbol{M}$. If $\boldsymbol{M}$ is of order N, and provided that the last row and column of $\boldsymbol{M}$ are left in last position during any reordering of rows and columns to make $\boldsymbol{L}$ more sparse, then $f(\boldsymbol{L}) = -\frac{1}{2}L_{NN}^2 - \sum_{i<N}\log(L_{ii})$ where L_{ij} is the ijth element of $\boldsymbol{L}$ and the summation applies only for nonzero L_{ii}.

Likelihood (11) is also a function of "$-\frac{1}{2}\log|\boldsymbol{R}| - \frac{1}{2}\log|\boldsymbol{G}|$". Because $\boldsymbol{R} = \sigma_e^2\boldsymbol{I}$ in most situations, $-\frac{1}{2}\log|\boldsymbol{R}|$ is easy to treat. However, $-\frac{1}{2}\log|\boldsymbol{G}|$ poses a greater challenge when (2) applies. There are simplifications that can be used with Toeplitz matrices (Zimmerman, 1989b), and can involve the Cholesky decomposition (Dietrich, 1993). Alternatively, $\boldsymbol{G}^{-1}$ can be subjected to sparse-matrix Cholesky factorization, noting that $-\frac{1}{2}\log|\boldsymbol{G}| = \frac{1}{2}\log|\boldsymbol{G}^{-1}|$. The relevant function is $f(\boldsymbol{L}) = \frac{1}{2}\log|\boldsymbol{G}^{-1}| = \sum_i \log(L_{ii})$ where $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{G}^{-1}$.

To facilitate maximum likelihood estimation of the dispersion parameters, Smith (1995) described how to differentiate $f(\boldsymbol{L})$, where $\boldsymbol{L}$ is computed directly via the Cholesky algorithm. These approaches describe the Cholesky algorithm as a collection of recursive steps where differentiation is applied directly as a forward progression through the recursions, or in a backward sweep. Whereas *forward differentiation* is preferable if there are several functions of $\boldsymbol{L}$ (rather than one) to differentiate, *backward differentiation* is faster when many different partial derivatives of a single function are needed (Griewank, 1989).

During the reverse sweep of backward differentiation, partial derivatives are accumulated using the chain rule of calculus. These partials are eventually related back to the initial parameters in the function. The end result is the evaluation of all first derivatives with one reverse pass through the recursion list. Multiple passes are needed for second derivatives. An accounting of the effort involved with backward differentiation of $f(\boldsymbol{L})$ is given in Table 1.

Once derivatives have been calculated, maximum likelihood estimation can proceed by a number of possible techniques. Some examples are the EM algorithm (e.g., Misztal and Perez-Enciso, 1993), the method of steepest ascent, and Fisher's method of scoring. The average-information method (Gilmour et al., 1995), like the EM algorithm, avoids the explicit calculation of second derivatives. But exact second derivatives are very computable due to the fore mentioned developments regarding automatic differentiation.

With $\boldsymbol{R}$ and $\boldsymbol{G}$ estimated, several interesting analyses are possible, including the conditional simulation; a popular tool for predicting the behavior of unknowns, or

Table 1
Effort[a] required to implement backward differentiation[b] on functions of the Cholesky decomposition, $f(\boldsymbol{L})$, where $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{C}$

Operation	Floating point operations	Floating point storage
$\boldsymbol{L}$	T[c]	S[d]
$\boldsymbol{L}$, first derivatives of $f(\boldsymbol{L})$	$3T$	$2S$
$\boldsymbol{L}$, first & second derivatives of $f(\boldsymbol{L})$	$(3 + 6n)T$	$4S$

[a] Ignores overhead cost for sparse matrix storage.
[b] Differentiate $f(\boldsymbol{L})$ with respect to n parameters.
[c] For dense matrices $T = N^3/6$, where N is the order of $\boldsymbol{C}$. In general, $T = \sum_i m_i^2/2$ where m_i is the number of nonzero elements of the ith column of $\boldsymbol{L}$.
[d] For dense matrices $S = N^2/2$. In general, S is the number of nonzero elements of $\boldsymbol{L}$.

missing data, when there is auxiliary information available for use in prediction (Gotway, 1994).

4.3. Conditional simulation

Not surprisingly, the Cholesky decomposition is also applicable for the conditional simulation of $\boldsymbol{b}$ when the definition (8) applies, and a frugal algorithm is outlined below:

(a) Find the Cholesky decomposition, $\boldsymbol{L}$, such that $\boldsymbol{LL}' = \boldsymbol{C}$.

(b) Evaluate $\boldsymbol{g} = \boldsymbol{L}^{-1}\boldsymbol{r}$ using forward substitution.

(c) Simulate a list ($\boldsymbol{n}$) of random normal zero-one deviates and add them to $\boldsymbol{g}$ to get $\hat{\boldsymbol{g}} = \boldsymbol{g} + \boldsymbol{n}$.

(d) To simulate one realization of $\boldsymbol{b}$, apply backward substitution to $\hat{\boldsymbol{g}}$ to get $\boldsymbol{L}^{-1'}\hat{\boldsymbol{g}}$.

Steps (c) and (d) can be repeated many times to simulate different realizations of $\boldsymbol{b}$. Linear combinations of $\boldsymbol{b}$ (that is, $\boldsymbol{Hb}$) will have the appropriate mean and variance (namely $\boldsymbol{H}\hat{\boldsymbol{b}}$ and $\boldsymbol{HC}^{-1}\boldsymbol{H}'$). Random residuals can also be simulated and added to $\boldsymbol{Hb}$ to obtain predictions of future observations or missing data.

The Cholesky decomposition and backward differentiation are also useful for crossvalidation, as found in the following illustration.

5. Nonparametric regression using a lattice

Simonoff (1996, p. 148) presents data consisting of grape yield for 52 rows of vineyard, and describes some standard nonparametric regression fits of yield to row number. Typical examples of nonparametric regression may involve piece-wise polynomials or local regression. But in this section a new nonparametric approach is presented that is based on a one-dimensional lattice, and it is demonstrated with the vineyard example.

While an interval on the x-axis is used to map the 52 rows, the interval is partitioned by 207 equally spaced subintervals that connect 208 elements. The mesh is fine enough to permit the approximation of a smooth function at each node. This means that 156 nodes correspond to missing data, a fact that causes no technical difficulty.

The statistical model is represented by

$$y_r = u_{x(r)} + e_r,$$

where y_r is the yield for the rth row, $u_{x(r)}$ is a smooth function (twice differentiable) evaluated at the coordinate $x(r)$, and e_r is the rth residual. To be consistent with Section 3.1, the model can be represented in matrix form as

$$\boldsymbol{y} = \boldsymbol{Zu} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a vector containing the yields, $\boldsymbol{u}$ a vector containing the 208 assignments of u_x, $\boldsymbol{e}$ is a vector containing the residuals, and $\boldsymbol{Z}$ is an incidence matrix that assigns elements of $\boldsymbol{u}$ to observations.

The residuals are assumed to be normally distributed, but because $\boldsymbol{u}$ has length 208, the statistical problem is highly under-determined. A smoothness restriction is needed, and what is usually advised is to add a special penalty function to the log-likelihood. The penalty involves integrating the square of the second derivatives of u_x, as indicated below:

$$-\tfrac{1}{2}\alpha\int[\partial^2 u_x/\partial x^2]^2\mathrm{d}x \approx -\tfrac{1}{2}\alpha\boldsymbol{u}'\boldsymbol{Q}\boldsymbol{u},$$

where α is the smoothness parameter, and $\boldsymbol{Q}$ is a banded matrix (penta-diagonal) and approximated by finite differences, i.e.,

$$\boldsymbol{Q} = \Delta'_{208\times 207}\Delta'_{207\times 206}\Delta_{206\times 207}\Delta_{207\times 208}$$

such that

$$\Delta = \begin{bmatrix} -1 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -1 & 1 & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ . & . & . & & & \\ 0 & 0 & 0 & & -1 & 1 \end{bmatrix}.$$

Therefore, the log-likelihood augmented by the penalty is approximately

$$-\tfrac{1}{2}\sigma^{-2}(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{u})'(\boldsymbol{y}-\boldsymbol{Z}\boldsymbol{u}) - \tfrac{1}{2}\alpha\boldsymbol{u}'\boldsymbol{Q}\boldsymbol{u},$$

where σ^2 is the variance of the residuals. To minimize this function, set the first derivatives to zero. This results in the mixed model equations with $\boldsymbol{Q}$ substituting for $\boldsymbol{G}^{-1}$:

$$[\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{Q}]\,\hat{\boldsymbol{u}} = \boldsymbol{Z}'\boldsymbol{y}, \quad \lambda = \alpha\sigma^2. \tag{12}$$

The matrix $\boldsymbol{Q}$ is rank deficient (rank = 206), and this accounts for the fact that the statistical model needs no intercept. Like $\boldsymbol{Q}$, the coefficient matrix is banded because $\boldsymbol{Z}'\boldsymbol{Z}$ is diagonal (or banded) when $\boldsymbol{Z}$ is a simple incidence matrix (or when $\boldsymbol{Z}$ represents a collection of interpolation formulas). Hence, with λ available it is very easy to estimate u_x. Most of the technical details are in the estimation of λ, however.

One way to estimate λ is by cross validation, i.e., by minimizing, PRESS, the error sums-of-squares resulting from predicting observations after they have been deleted from the data in turn. Note that if observations are not removed, the resulting fit would be stupidly perfect with small λ. The criterion for minimization is, therefore,

$$\mathrm{PRESS} = \tfrac{1}{2}\sum_r (y_r - \boldsymbol{z}_r'\boldsymbol{u}_r)^2,$$

where z_r' is the rth row of $\boldsymbol{Z}$, and u_r is the estimate of $\boldsymbol{u}$ after deleting the rth observation:

$$\boldsymbol{u}_r = [\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{Q} - z_r z_r']^{-1}\,[\boldsymbol{Z}'\boldsymbol{y} - y_r \cdot z_r].$$

Even with a banded coefficient matrix, the calculation of PRESS would appear formidable because of the outer iteration involving the row index r. However, Cook (1977) noted that PRESS can be simplified by using an identity to re-express the inverse of a matrix sum:

$$\text{PRESS} = \frac{1}{2}\sum_r \left[\frac{y_r - z_r'\hat{\boldsymbol{u}}}{1 - h_r}\right]^2.$$

where $\hat{\boldsymbol{u}}$ solves (12) and h_r is the leverage, $z_r'(\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{Q})^{-1}z_r$. While this equation provides an important simplification, it still implies the calculation of h_r for each r. But coincidentally an important relationship is uncovered by introducing the diagonal weight matrix $\text{diag}\{w_r\} = \boldsymbol{W}$, and observing that $h_r = \partial \log|\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{Q}|/\partial w_r$ when $\boldsymbol{W} = \boldsymbol{I}$. In other words, the leverage statistics are merely derivatives of the log-determinant of the coefficient matrix, and hence they can all be computed jointly in one reverse pass through the Cholesky decomposition using backward differentiation. However, in order to minimize PRESS it is still desirable to have the gradient of PRESS and perhaps some approximate second derivatives. Interestingly, these can also be obtained in linear time, with no double iteration involving r. There are several ways to do this, and one approach is outlined below:

(1) Compute the Cholesky decomposition $\boldsymbol{L}\boldsymbol{L}' = \boldsymbol{C}$, where $\boldsymbol{C} = \boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z} + \lambda\boldsymbol{Q}$ and $\boldsymbol{W} = \boldsymbol{I}$.

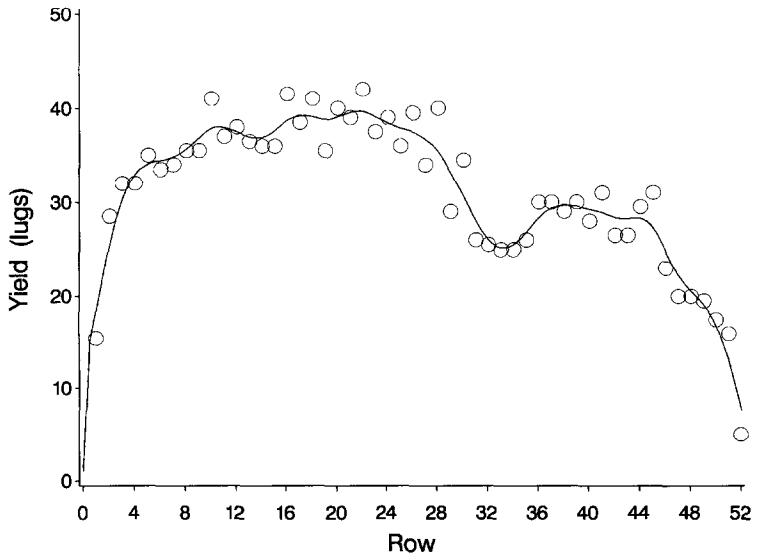

Fig. 2. Smooth function fitted to vineyard data.

(2) Evaluate $h_r = \partial \log|\boldsymbol{C}|/\partial w_r$ (all r), by backward differentiation applied to the Cholesky algorithm (Smith, 1995).

(3) Apply Smith's (1995) algorithm for second derivatives to compute $\partial h_r/\partial\lambda = \partial^2 \log|\boldsymbol{C}|/\partial\lambda\partial w_r$ (all r), and $\partial \boldsymbol{L}/\partial\lambda$.

(4) Apply forwards and backward substitution to get $\hat{\boldsymbol{u}}$, and with these steps compute $\partial\hat{\boldsymbol{u}}/\partial\lambda$ using simple forward differentiation given $\partial \boldsymbol{L}/\partial\lambda$.

(5) Assemble the various items involving $\hat{e}_r = (y_r - \boldsymbol{z}_r'\hat{\boldsymbol{u}})/(1 - h_r)$:

a. PRESS $= \frac{1}{2}\sum_r \hat{e}_r^2$;

b. $\partial \text{PRESS}/\partial\lambda = \sum_r \hat{e}_r \cdot \partial\hat{e}_r/\partial\lambda$;

c. An approximation $\partial^2 \text{PRESS}/\partial\lambda^2 \approx \sum_r [\partial\hat{e}_r/\partial\lambda]^2$.

This set of instructions permits a quasi-Newton minimization of PRESS in real time. In the vineyard example, two minima were identified with this algorithm, and the fit ($\lambda = 89.9$) with the smallest PRESS is presented in Fig. 2. The alternative fit was smoother ($\lambda = 612.9$) and possessed only a slightly larger PRESS. Selecting a fit from the two requires some professional opinion.

6. Remarks

This paper shows that sparse-matrix tools can be very useful for analyses involving lattices. The work required is related to the fill-in that occurs during the Cholesky decomposition, and this interacts with the dimension of the lattice. In general, two-dimensional lattices require more work than the one-dimensional lattice described in Section 5, and three dimensions requires more work than two. However, the advantage of sparse-matrix methods relative to some kriging examples is that the work no longer grows as the cube of the number of observations, and all missing observations are computed jointly by solving one linear system. Like kriging, the tools may accommodate irregularly spaced data as a close approximation if the mesh is sufficiently fine. Some of the more classical techniques developed for lattices do not treat missing values.

References

Basu, S., Reinsel, G.C., 1993. Properties of the spatial unilateral first-order ARMA model. Adv. Appl. Probab. 25, 631–648.

Basu, S., Reinsel, C.G., 1994. Regression models with spatially correlated errors. J. Amer. Statist. Assoc., 89 (425), 88–99.

Besag, J.E., 1974. Spatial interaction and the statistical analysis of lattice systems. J. Roy. Statist. Soc. B 36, 192–225.

Cook, R.D., 1977. Detection of Influential Observations in Linear Regression. Technometrics 19, 15–18.

Cullis, B.R., Gleeson, A.C., 1991. Spatial analysis of field experiments – an extension to two dimensions. Biometrics 47, 1449–1460.

Cressie, N., 1991. Statistics for Spatial Data. Wiley, New York.

Dempfle, L., 1977. Relation entre BLUP (best linear unbiased prediction) et estimateurs Bayesiens, Ann Génétique et de Sélection Animale 9, 27–32.

Dietrich, C.R., 1993. Computationally efficient Cholesky factorization of a covariance matrix with block Toeplitz structure. J. Statist. Comput. Simulation 45, 203–218.

George, A., J.W.H. Liu, 1981. Computer Solutions of Large Sparse Positive Definite Systems. Prentice-Hall, Englewood Cliffs, NJ.

Gilmour, A.R., Thompson, R., Cullis, B.R., 1995. Average information REML: an efficient algorithm for variance parameter estimation in linear mixed models. Biometrics 51, 1440–1450.

Gleeson, A.C., Cullis, B.R., 1987. Residual maximum likelihood (REML) estimation of a neighbor model for field experiments. Biometrics 43, 277–288.

Gleeson, A.C., McGilchrist, C.A., 1980. Bilateral processes on a rectangular lattice. Austral. J. Statist. 22 (2), 197–206.

Goldberger, A.S., 1962. Best linear unbiased prediction in the generalized linear regression model. J. Amer. Statist. Assoc., 70, 369–375.

Golub, G.H., Van Loan, C.F., 1983. Matrix Computation. The John Hopkins University Press, Baltimore, MD.

Gotway, C.A., 1994. The use of conditional simulation in nuclear-waste-site performance assessment, with discussions. Technometrics 36 (2), 129–161.

Graser, H.-U., Smith, S.P., Tier, B., 1987. A derivative free approach for estimating variance components in animal models by REML. J. Animal Sci. 64, 1362–1370.

Griewank, A., 1989. On automatic differentiation, in mathematical programming: recent developments and applications. In: Iri, M., Tanabe, K. (Eds.), Mathematical Programming. KTK Scientific Publishers, Tokyo, pp. 83–107.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ.

Harville, D.A., 1976. Extension of the Gauss–Markov theorem to include the estimation of random effects. Ann. Statist. 4, 384–395.

Harville, D.A., 1977. Maximum likelihood approaches to variance component estimation and to related problems. J. Amer. Statist. Assoc. 72 (358), 320–340.

Henderson, C.R., Kempthorne, O., Searle, S.R., Von Krosigk, C.N., 1959. Estimation of environmental and genetic trends from records subject to culling. Biometrics 15, 192–218.

Henderson, H.V., Searle, S.R., 1981. On deriving the inverse of a sum of matrices. SIAM Rev. 23 (1), 53–60.

Hudson, G.F.S., 1986. Computing genetic evaluations through application of generalized least squares to an animal model. Génétique Sélection Évolution 18, 31–40.

Jones, R.H., Vecchia, A.V., 1993. Fitting continuous ARMA models to unequally spaced spatial data. J. Amer. Statist. Assoc., 88 (423), 947–954.

Laslett, G.M., 1994. Kriging and splines: an empirical comparison of their predictive performance in some applications. J. Amer. Statist. Assoc. 89 (426), 391–400.

Martin, R., 1979. A subclass of lattice processes applied to a problem in planar sampling. Biometrika, 66, 209–217.

Mead, R., 1967. A mathematical model for the estimation of inter-plant competition. Biometrics 23, 659–671.

Misztal, I., Perez-Enciso, M., 1993. Sparse matrix inversion for restricted maximum likelihood estimation of variance components by expectation – Maximization. J. Dairy Sci., 76 (5), 1479–1483.

Ng, E., Peyton, B.W., 1993a. A supernodal Cholesky factorization algorithm for shared-memory multiprocessors. SIAM J. Sci. Comput. 14 (4), 761–769.

Ng, E.G., Peyton, B.W., 1993b. Block sparse Cholesky algorithms on advanced uniprocessor computers. SIAM J. Sci. Comput. 14 (5), 1034–1056.

Patterson, H.D., Thompson, R., 1971. Recovery of inter-block information when block sizes are unequal. Biometrika 58, 545–554.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6, 15–51.

Rothberg, E., Gupta, A., 1994. An efficient block-oriented approach to sparse Cholesky factorization. SIAM J. Sci. Comput. 15 (6), 1413–1439.

Simonoff, J.S., 1996. Smoothing Methods in Statistics. Springer, New York.

Smith, S.P., 1995. Differentiation of the Cholesky algorithm. J. Comput. Graph. Statist., 4 (2), 134–147.

Smith, S.P., 1997. Kalman filtering with the Cholesky decomposition. J. Comput. Graph. Statist., submitted.

Vecchia, A.V., 1992. A method of prediction for spatial regression models with correlated errors. J. Roy. Statist. Soc. Ser. B 54 (3), 813–830.

Whittle, P., 1954. On stationary processes in the plane. Biometrika 41, 434–449.

Zimmerman, D.L., 1989a. Computationally efficient restricted maximum likelihood estimation of generalized covariance functions. Math. Geology 2 (7), 655–672.

Zimmerman, D.L., 1989b. Computationally exploitable structure of covariance matrices and generalized covariance matrices in spatial models. J. Statist. Comput. Simulation 32, 1–15.

Zimmerman, D.L., Harville, D.A., 1991. A random field approach to the analysis of field-plot experiments. Biometrics 47, 223–239.