Statistics and Computing (1996) **6**, 289–296

Unconstrained parametrizations for variance–covariance matrices

JOSÉ C. PINHEIRO[1] and DOUGLAS M. BATES[2]

[1]Department of Biostatistics and [2]Department of Statistics, University of Wisconsin–Madison, Madison, Wisconsin, USA

Submitted January 1995 and accepted August 1995

The estimation of variance–covariance matrices through optimization of an objective function, such as a log-likelihood function, is usually a difficult numerical problem. Since the estimates should be positive semi-definite matrices, we must use constrained optimization, or employ a parametrization that enforces this condition. We describe here five different parametrizations for variance–covariance matrices that ensure positive definiteness, thus leaving the estimation problem unconstrained. We compare the parametrizations based on their computational efficiency and statistical interpretability. The results described here are particularly useful in maximum likelihood and restricted maximum likelihood estimation in linear and non-linear mixed-effects models, but are also applicable to other areas of statistics.

Keywords: Unconstrained estimation, variance–covariance components estimation, Cholesky factorization, matrix logarithm

1. Introduction

The estimation of variance–covariance matrices through optimization of an objective function, such as a log-likelihood function, is usually a difficult numerical problem, since one must ensure that the resulting estimate is positive semi-definite. This kind of estimation problem occurs, for example, in the analysis of linear and non-linear mixed-effects models, when one is interested in obtaining maximum likelihood, or restricted maximum likelihood, estimates of the random effects variance–covariance matrix (Lindstrom and Bates, 1988, 1990). In these models, because the random effects are unobserved quantities, no *sample variance–covariance matrix* type of estimator, that would automatically be positive semi-definite, is available. Indirect estimation methods must be used. The estimation of a variance–covariance matrix through optimization of a log-likelihood function may occur even when a *sample variance–covariance* estimator is available, if, for example, one is interested in using maximum likelihood asymptotic results to assess the variability in the resulting estimates.

Two approaches can be used to ensure positive semi-definiteness of a variance–covariance matrix estimate: constrained optimization, where the natural parametrization for the upper-triangular elements in the variance–covariance matrix is used and the estimates are constrained to be positive semi-definite matrices; and unconstrained optimization, where the upper-triangular elements in the variance–covariance matrix are reparametrized in such a way that the resulting estimate must be positive semi-definite.

The first approach, constrained optimization using the non-redundant entries of the matrix Σ as parameters, would be very difficult. As Dennis and Schnabel (1983) point out, attempts to solve a constrained optimization problem usually boil down to repeated unconstrained problems or to solving a non-linear system of equations. The simplest cases are those where there are simple inequality constraints on the parameters and even in those cases constrained solutions can require several times as much effort as an unconstrained solution. In addition, the statistical properties of constrained estimates, such as asymptotic properties, can be difficult to characterize.

In this case, the constraints themselves are quite complicated to express. Simply verifying that a given symmetric matrix is positive semi-definite is at least as difficult as calculating the matrix's Cholesky factorization, which we will describe later. Most of the parametrizations we will describe require less effort than this at each iteration in transforming from the given values of the parameters to the matrix $\boldsymbol{\Sigma}$.

For these reasons, we recommend the use of unconstrained optimization with a parametrization that enforces the positive semi-definite constraint.

An unconstrained estimation approach for variance–covariance matrices in a Bayesian context using matrix logarithms can be found in Leonard and Hsu (1993). Lindstrom and Bates (1988, 1990) describe the use of Cholesky factors for implementing unconstrained estimation of random effects variance–covariance matrices in linear and non-linear mixed effects models using likelihood and restricted likelihood.

Since a variance–covariance matrix is positive semi-definite, but not positive definite, only in the rather degenerate situation of linear combinations of the underlying random variables taking constant values, we will restrict ourselves here to positive definite variance–covariance matrices.

In addition to enforcing the positive definiteness constraints, the choice of the parametrization can be influenced by computational efficiency and by the statistical interpretability of the individual elements. In general, we can use numerically or analytically determined second derivatives of the objective function to approximate standard errors and derive confidence intervals for the individual parameters. To assess the variability of the variances and covariances estimates, it is desirable that they can be expressed as simple functions of the unconstrained parameters. More detailed techniques, such as profiling the likelihood (Bates and Watts, 1988, Chapter 6), also work best for functions of the variance–covariance matrix that are expressed in the original parametrization.

We describe in Section 2 five different parametrizations for transforming the estimation of unstructured (general) variance–covariance matrices into an unconstrained problem. In Section 3 we compare the parametrizations with respect to their computational efficiency and statistical interpretability. Our conclusions and suggestions for further research are presented in Section 4.

2. Parametrizations

Let $\boldsymbol{\Sigma}$ denote a symmetric positive definite $n \times n$ variance–covariance matrix corresponding to a random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$. We do not assume any further structure for $\boldsymbol{\Sigma}$. Because $\boldsymbol{\Sigma}$ is symmetric, only $n(n+1)/2$ parameters are needed to represent it. We will denote by $\boldsymbol{\theta}$ any such minimal set of parameters to determine $\boldsymbol{\Sigma}$. The rationale behind all parametrizations considered in this section is to write

$$\boldsymbol{\Sigma} = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L} \tag{1}$$

where $\boldsymbol{L} = \boldsymbol{L}(\boldsymbol{\theta})$ is an $n \times n$ matrix of full rank obtained from a $n(n+1)/2$-dimensional vector of unconstrained parameters $\boldsymbol{\theta}$. It is clear that any $\boldsymbol{\Sigma}$ defined from a full rank $\boldsymbol{L}$ as in (1) is positive definite.

Different choices of $\boldsymbol{L}$ lead to different parametrizations of $\boldsymbol{\Sigma}$. We will consider here two classes of $\boldsymbol{L}$: one based on the Cholesky factorization (Thisted, 1988, §3.3) of $\boldsymbol{\Sigma}$ and another based on the spectral decomposition of $\boldsymbol{\Sigma}$ (Rao, 1973, §1c.3). The first three parametrizations presented below use the Cholesky factorization of $\boldsymbol{\Sigma}$, while the last two are based on its spectral decomposition.

In some of the parametrizations there are particular components of the parameter vector $\boldsymbol{\theta}$ that have meaningful statistical interpretations. These can include the eigenvalues of $\boldsymbol{\Sigma}$, which are important in considering when the matrix is ill-conditioned, the individual variances or standard deviations, and the correlations.

The following variance–covariance matrix will be used throughout this section to illustrate the use of the various parametrizations:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 5 & 5 \\ 1 & 5 & 14 \end{bmatrix} \tag{2}$$

2.1. *Cholesky parametrization*

Because $\boldsymbol{\Sigma}$ is positive definite, it may be factored as $\boldsymbol{\Sigma} = \boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}$, where $\boldsymbol{L}$ is an upper triangular matrix. Setting $\boldsymbol{\theta}$ to be the upper triangular elements of $\boldsymbol{L}$ gives the Cholesky parametrization of $\boldsymbol{\Sigma}$. Lindstrom and Bates (1988) use this parametrization to obtain derivatives of the log-likelihood of a linear mixed-effects model for use in a Newton–Raphson algorithm. They reported that the use of this parametrization dramatically improved the convergence properties of the optimization algorithm, when compared to a constrained estimation approach.

One problem with the Cholesky parametrization is that the Cholesky factor is not unique. In fact, if $\boldsymbol{L}$ is a Cholesky factor of $\boldsymbol{\Sigma}$ then so is any matrix obtained by multiplying a subset of the rows of $\boldsymbol{L}$ by -1. This has implications on parameter identification, since up to 2^n different $\boldsymbol{\theta}$ may represent the same $\boldsymbol{\Sigma}$. Numerical problems can arise in the optimization of an objective function when different optimal solutions are close together in the parameter space.

Another problem with the Cholesky parametrization is the lack of a straightforward relationship between $\boldsymbol{\theta}$ and the elements of $\boldsymbol{\Sigma}$. This makes it hard to interpret the estimates of $\boldsymbol{\theta}$ and to obtain confidence intervals for the variances and covariances in $\boldsymbol{\Sigma}$ based on confidence intervals for the estimates of $\boldsymbol{\theta}$. One exception is $|[\boldsymbol{L}]_{11}| = \sqrt{[\boldsymbol{\Sigma}]_{11}}$, so confidence intervals on $[\boldsymbol{\Sigma}]_{11}$ can be obtained from confidence intervals on $[\boldsymbol{L}]_{11}$, where $[\boldsymbol{A}]_{ij}$ denotes the ijth element of the matrix $\boldsymbol{A}$. By appropriately permuting the columns and rows of $\boldsymbol{\Sigma}$ we can in fact derive confidence intervals for all the variance terms based on confidence intervals for the elements of $\boldsymbol{L}$.

The main advantage of this parametrization, apart from the fact that it ensures positive definiteness of the estimate of $\boldsymbol{\Sigma}$, is that it is computationally simple and stable.

The Cholesky factorization of $\boldsymbol{A}$ in (2) is

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

By convention, the components of the upper triangular part of $\boldsymbol{L}$ are listed columnwise to give $\boldsymbol{\theta} = (1, 1, 2, 1, 2, 3)^{\mathrm{T}}$.

2.2. *Log-Cholesky parametrization*

If one requires the diagonal elements of $\boldsymbol{L}$ in the Cholesky factorization to be positive then $\boldsymbol{L}$ is unique. To avoid constrained estimation, one can use the logarithms of the diagonal elements of $\boldsymbol{L}$. We call this parametrization the *log-Cholesky parametrization*. It inherits the good computational properties of the Cholesky parametrization, but has the advantage of being uniquely defined. As in the Cholesky parametrization the parameters lack direct interpretation in terms of the original variances and covariances, except for L_{11}.

The log-Cholesky parametrization of $\boldsymbol{A}$ is $\boldsymbol{\theta} = (0, 1, \log(2), 1, 2, \log(3))^{\mathrm{T}}$.

2.3. *Spherical parametrization*

The purpose of this parametrization is to combine the computational efficiency of the Cholesky parametrization with direct interpretation of $\boldsymbol{\theta}$ in terms of the variances and correlations in $\boldsymbol{\Sigma}$.

Let $\boldsymbol{L}_i$ denote the ith column of $\boldsymbol{L}$ in the Cholesky factorization of $\boldsymbol{\Sigma}$ and $\boldsymbol{l}_i$ denote the spherical coordinates of the first i elements of $\boldsymbol{L}_i$. That is,

$$\begin{aligned} [\boldsymbol{L}_i]_1 &= [\boldsymbol{l}_i]_1 \cos([\boldsymbol{l}_i]_2) \\ [\boldsymbol{L}_i]_2 &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \cos([\boldsymbol{l}_i]_3) \\ &\dots \\ [\boldsymbol{L}_i]_{i-1} &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \cdots \cos([\boldsymbol{l}_i]_i) \\ [\boldsymbol{L}_i]_i &= [\boldsymbol{l}_i]_1 \sin([\boldsymbol{l}_i]_2) \cdots \sin([\boldsymbol{l}_i]_i) \end{aligned}$$

It then follows that $\Sigma_{ii} = [\boldsymbol{l}_i]_1^2$ and $\rho_{1i} = \cos([\boldsymbol{l}_i]_2)$, $i = 2, \dots, n$, where ρ_{ij} denotes the correlation coefficient between X_i and X_j. The correlations between other variables can be expressed as linear combinations of products of sines and cosines of the elements in $\boldsymbol{l}_1, \dots, \boldsymbol{l}_n$, but the relationship is not as straightforward as those involving X_1. If confidence intervals are available for the elements of $\boldsymbol{l}_i$, $i = 1, \dots, n$ then we can also obtain confidence intervals for the variances and the correlations ρ_{1i}. By appropriately permuting the rows and columns of $\boldsymbol{\Sigma}$, we can in fact obtain confidence intervals for all the variances and correlations of $X_1, \dots, X_n$. The exact same reasoning can be applied to derive profile traces and profile contours (Bates and Watts, 1988) for variances and correlations of $X_1, \dots, X_n$ based on a likelihood function.

To ensure uniqueness of the spherical parametrization we must have

$$\begin{aligned} &[\boldsymbol{l}_i]_1 > 0, && i = 1, \dots, n \quad \text{and} \\ &[\boldsymbol{l}_i]_j \in (0, \pi), && i = 2, \dots, n, \quad j = 2, \dots, i \end{aligned}$$

Unconstrained estimation is obtained by defining $\boldsymbol{\theta}$ as follows:

$$\theta_i = \log([[\boldsymbol{l}_i]_1), \quad i = 1, \dots, n \qquad \text{and}$$

$$\theta_{n+(i-2)(i-1)/2+(j-1)} = \log\left(\frac{[\boldsymbol{l}_i]_j}{\pi - [\boldsymbol{l}_i]_j}\right),$$

$$i = 2, \dots, n, \; j = 2, \dots, i.$$

The spherical parametrization has about the same computational efficiency as the Cholesky and log-Cholesky parametrizations, is uniquely defined, and allows direct interpretability of $\boldsymbol{\theta}$ in terms of the variances and correlations in $\boldsymbol{\Sigma}$.

The spherical parametrization of $\boldsymbol{A}$ is $\boldsymbol{\theta} = (0, \log(5)/2, \log(14)/2, -0.608, -0.348, -0.787)^{\mathrm{T}}$.

2.4. Matrix logarithm parametrization

This and the next parametrization are based on the spectral decomposition of $\boldsymbol{\Sigma}$. Because $\boldsymbol{\Sigma}$ is positive definite, it has n positive eigenvalues $\boldsymbol{\lambda}$. Letting $\boldsymbol{U}$ denote the orthogonal matrix of orthonormal eigenvectors of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$, we can write

$$\boldsymbol{\Sigma} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\mathrm{T}} \tag{3}$$

By setting

$$\boldsymbol{L} = \boldsymbol{\Lambda}^{1/2}\boldsymbol{U}^{\mathrm{T}} \tag{4}$$

in (1), where $\boldsymbol{\Lambda}^{1/2}$ denotes the diagonal matrix with $[\Lambda^{1/2}]_{ii} = \sqrt{[\Lambda]_{ii}}$, we get a factorization of $\boldsymbol{\Sigma}$ based on the spectral decomposition.

The matrix logarithm of $\boldsymbol{\Sigma}$ is defined as $\log(\boldsymbol{\Sigma}) = \boldsymbol{U}\log(\boldsymbol{\Lambda})\boldsymbol{U}^{\mathrm{T}}$, where $\log(\boldsymbol{\Lambda}) = \operatorname{diag}[\log(\boldsymbol{\lambda})]$. Note that $\boldsymbol{\Sigma}$ and $\log(\boldsymbol{\Sigma})$ share the same eigenvectors. The matrix $\log(\boldsymbol{\Sigma})$ can take any value in the space of $n \times n$ symmetric matrices. Letting $\boldsymbol{\theta}$ be equal to its upper triangular elements gives the matrix logarithm parametrization of $\boldsymbol{\Sigma}$.

The matrix logarithm parametrization defines a one-to-one mapping between $\boldsymbol{\theta}$ and $\boldsymbol{\Sigma}$ and therefore does not have the identification problems of the Cholesky factorization. It does involve considerable calculations, as $\boldsymbol{\theta}$ produces $\log(\boldsymbol{\Sigma})$ whose eigenstructure must be determined before $\boldsymbol{L}$ in (4) can be calculated. Similarly to the Cholesky and log-Cholesky parametrizations, the vector $\boldsymbol{\theta}$ in the matrix logarithm parametrization does not have a straightforward interpretation in terms of the original variances and covariances in $\boldsymbol{\Sigma}$. We note that even though the matrix logarithm is based on the spectral decomposition of $\boldsymbol{\Sigma}$, there is not a straightforward relationship between $\boldsymbol{\theta}$ and the eigenvalues–eigenvectors of $\boldsymbol{\Sigma}$.

The matrix logarithm of $\boldsymbol{A}$ is

$$\log(\boldsymbol{A}) = \begin{bmatrix} -0.174 & 0.392 & 0.104 \\ 0.392 & 1.265 & 0.650 \\ 0.104 & 0.650 & 2.492 \end{bmatrix}$$

and therefore the matrix logarithm parametrization of $\boldsymbol{A}$ is $\boldsymbol{\theta} = (-0.174, 0.392, 1.265, 0.104, 0.650, 2.492)^{\mathrm{T}}$.

2.5. *Givens parametrization*

The eigenstructure of $\boldsymbol{\Sigma}$ contains valuable information for determining whether some linear combination of $X_1, \ldots, X_n$ could be regarded as *nearly* constant. This is useful, for example, in model building for mixed-effects models, as *near zero* eigenvalues may indicate overparametrization (Pinheiro and Bates, 1995). The Givens parametrization uses the eigenvalues of $\boldsymbol{\Sigma}$ directly in the definition of the parameter vector $\boldsymbol{\theta}$.

The Givens parametrization is based on the spectral decomposition of $\boldsymbol{\Sigma}$ given in (3) and the fact that the eigenvector matrix $\boldsymbol{U}$ can be represented by $n(n-1)/2$ angles, used to generate a series of Givens rotation matrices (Thisted, 1988, §3.1.6.6) whose product reproduce $\boldsymbol{U}$ as follows:

$$\boldsymbol{U} = \boldsymbol{G}_1\boldsymbol{G}_2 \cdots \boldsymbol{G}_{n(n-1)/2},$$

where

$$\boldsymbol{G}_i[j,k] = \begin{cases} \cos(\delta_i), & \text{if } j = k = m_1(i) \text{ or } j = k = m_2(i) \\ \sin(\delta_i), & \text{if } j = m_1(i), k = m_2(i) \\ -\sin(\delta_i), & \text{if } j = m_2(i), k = m_1(i) \\ 1, & \text{if } j = k \neq m_1(i) \text{ and } j = k \neq m_2(i) \\ 0, & \text{otherwise} \end{cases}$$

and $m_1(i) < m_2(i)$ are integers taking values in $\{1, \ldots, n\}$ and satisfying $i = m_2(i) - m_1(i) + (m_1(i) - 1)(n - m_1(i)/2)$. To ensure uniqueness of the Givens parametrization we must have $\delta_i \in (0, \pi)$, $i = 1, \ldots, n(n-1)/2$.

The spectral decomposition (3) is unique up to a reordering of the diagonal elements of $\boldsymbol{\Lambda}$ and columns of $\boldsymbol{U}$ and up to switching of signs in each column of $\boldsymbol{U}$. Uniqueness can be achieved by forcing the eigenvalues to be sorted in ascending order. This can be attained, within an unconstrained estimation framework, by using a parametrization suggested by Jupp (1978) and defining the first n elements of $\boldsymbol{\theta}$ as

$$\theta_i = \log(\lambda_i - \lambda_{i-1}), \qquad i = 1, \ldots, n$$

where λ_i denotes the ith eigenvalue of $\boldsymbol{\Sigma}$ is ascending order and with the convention that $\lambda_0 = 0$. The remaining elements of $\boldsymbol{\theta}$ in the Givens parametrization are defined by the relation

$$\theta_{n+i} = \log\left(\frac{\delta_i}{\pi - \delta_i}\right), \qquad i = 1, \ldots, n(n-1)/2.$$

The main advantage of this parametrization is that the first n elements of $\boldsymbol{\theta}$ give information about the eigenvalues of $\boldsymbol{\Sigma}$ directly. Another advantage of the Givens parametrization is that it can be easily modified to handle general (not necessarily positive definite) symmetric matrices. The only modification needed is to set $\theta_1 = \lambda_1$ and

$$\lambda_i = \theta_1 + \sum_{j=2}^{i} \exp(\theta_i), \qquad i = 2, \ldots, n.$$

The main disadvantage of this parametrization is that it involves considerable computational effort in the calculation of $\boldsymbol{\Sigma}$ from the parameter vector $\boldsymbol{\theta}$. Another problem with the Givens parametrization is that one cannot relate $\boldsymbol{\theta}$ to the elements of $\boldsymbol{\Sigma}$ in a straightforward manner, so inferences about variances and covariances require indirect methods.

The eigenvector matrix $\boldsymbol{U}$ in (3) can also be expressed as a product of a series of Householder reflection matrices (Thisted, 1988, §3.1.2) and these in turn can be derived from $n(n-1)/2$ parameters used to obtain the directions of the Householder reflections (Pinheiro, 1994). This Householder parametrization is essentially equivalent to the Givens parametrization in terms of statistical interpretability, but it is less efficient, since the derivation of the Householder reflection matrices involves even more computation than the Givens rotations. We have not considered it here.

The Givens parametrization of $\boldsymbol{A}$ is $\boldsymbol{\theta} = (-0.275, 0.761, 2.598, -0.265, -0.562, -0.072)^{\mathrm{T}}$.

3. Comparing the parametrizations

In this section we compare the parametrizations described in Section 2 in terms of their computational efficiency and the statistical interpretability of the individual parameters.

Table 1. *Different eigenvalue structures for n × n matrices* $\boldsymbol{\Sigma}$ *used in the simulation study*

Structure	Eigenvalues
I	$\{1, 1, \ldots, 1, 1\}$
II	$\{1000, 1, 1, \ldots, 1, 1\}$
III	$\{1, 1, \ldots, 1, 0.001\}$
IV	$\left\{\underbrace{1000, \ldots, 1000}_{n/2}, \underbrace{0.001, \ldots, 0.001}_{n/2}\right\}$
V	$\{1000, 1, \ldots, 1, 0.001\}$
VI	$\{10, 20, 30, \ldots, (n-1) \times 10, n \times 10\}$

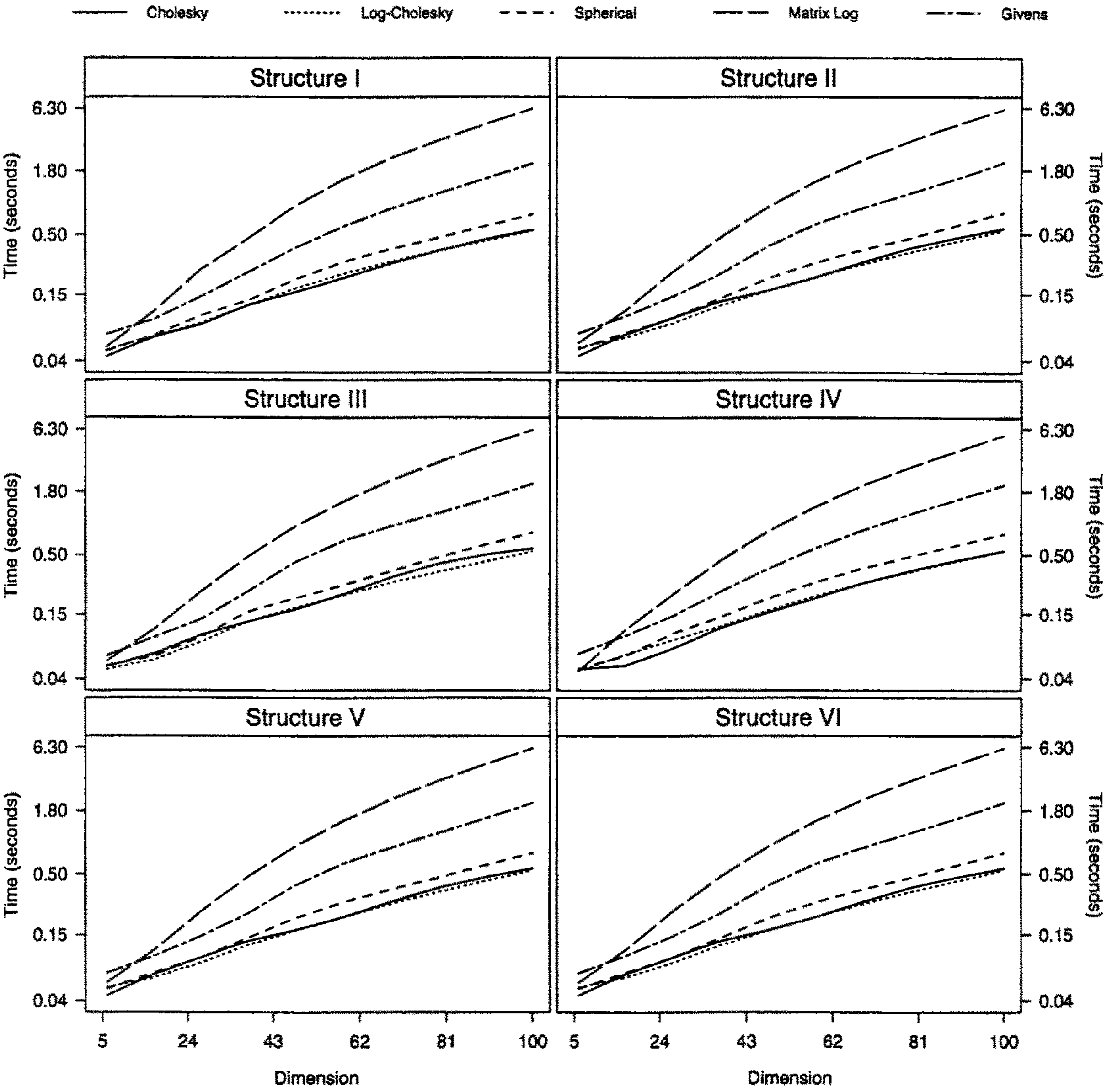

Fig. 1. *Average user time to calculate* $\boldsymbol{L}$ *as a function of* n, *for the different parametrizations and eigenstructures of* $\boldsymbol{\Sigma}$

The computational efficiency of the different parametrizations is assessed by simulation. First, we analyse the average time needed to calculate $\boldsymbol{L}(\boldsymbol{\theta})$ from $\boldsymbol{\theta}$ for each parametrization and for different eigenstructures and for varying sizes of $\boldsymbol{\Sigma}$. Then we compare the performance of the different parametrizations in computing the maximum likelihood estimate of the variance–covariance matrix in a linear mixed–effects model (Laird and Ware, 1982).

To investigate the effect of the eigenstructure of $\boldsymbol{\Sigma}$ on the computational efficiency of the parametrizations, six different eigenvalue structures, described in Table 1, were considered in the simulation study presented below.

Random $\boldsymbol{\Sigma}$ matrices of dimension n, for a given eigenvalue structure $(\lambda_1, \ldots, \lambda_n)$, were generated according to the following algorithm.

Step 1. Select a random n-dimensional orthogonal matrix $\boldsymbol{U}$ uniformly on the group of orthogonal matrices, using the algorithm proposed by Anderson *et al.* (1987).

Step 2. Generate n independent random variables $X_1, \ldots, X_n$, such that $X_i \sim \mathcal{N}(\log(\lambda_i), 0.01)$, and form a diagonal matrix of random eigenvalues, $\boldsymbol{\Lambda}$, with $[\boldsymbol{\Lambda}]_{ii} = \exp(X_i)$. This ensures that the relative variability of the random eigenvalues is the same.

Step 3. Obtain $\boldsymbol{\Sigma} = \boldsymbol{U\Lambda U}^{\mathrm{T}}$.

To evaluate the average time needed to calculate $\boldsymbol{L}$, we generated, for each of the eigenvalue structures in Table 1, 25 random $n \times n$ matrices $\boldsymbol{\Sigma}$ according to the above algorithm, with n varying from 6 to 100. For each $\boldsymbol{\Sigma}$ we obtained $\boldsymbol{\theta}$ and recorded the average time to calculate $\boldsymbol{L}$. The time quoted is the time the CPU spent

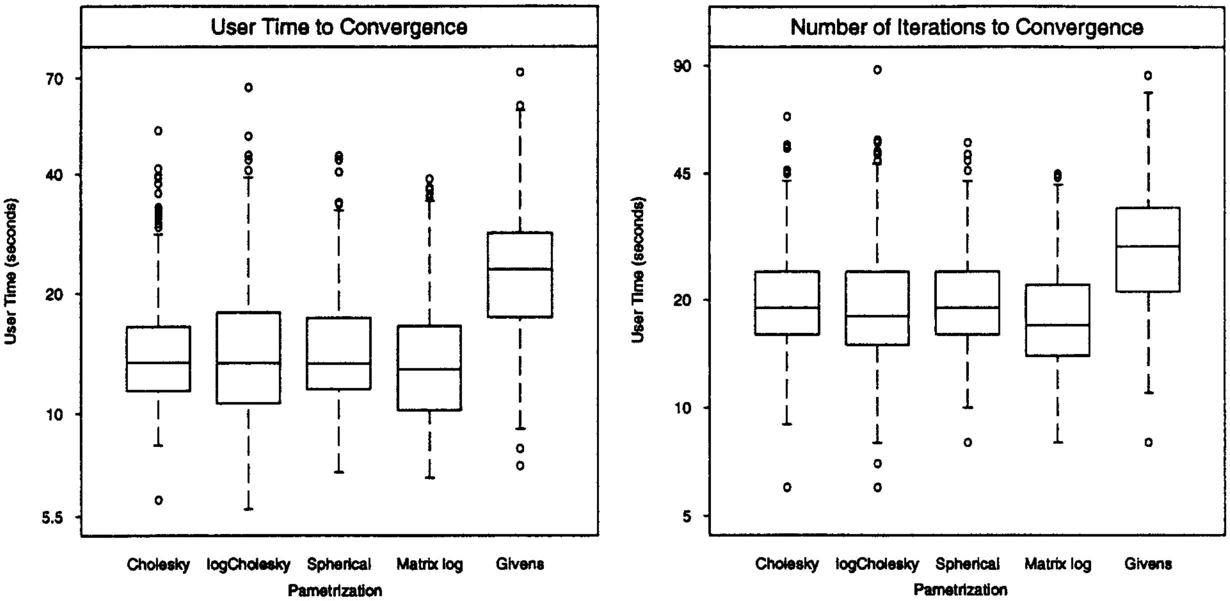

Fig. 2. *Box-plots of user time and number of iterations to convergence for 300 random samples of model (5) with Σ of dimension 3*

evaluating the user's program for the calculation. Because these *user* times were too small for accurate evaluation when using matrices of dimension less than 10, we used 5 evaluations of $\boldsymbol{L}$ for each user time calculation. Figure 1 presents the average user time as a function of n for each of the parametrizations of Σ and for each of the eigenvalue structures in Table 1.

The computational performances of the parametrizations are essentially the same for all eigenvalue structures considered. The Cholesky, the log-Cholesky, and the spherical parametrizations have similar performances, considerably better than the other two parametrizations. The Cholesky and the log-Cholesky parametrizations have slightly better performances than the spherical parametrization, especially for $n \geq 25$. The matrix logarithm had the worst performance, followed by the Givens parametrization. These results are essentially reflecting the computational complexity of each parametrization, as described in Section 2.

To compare the different parametrizations in an estimation context, we conducted a small simulation study using the linear mixed effects model

$$y_i = X_i(\beta + b_i) + \varepsilon_i, \qquad i = 1, \ldots, M \tag{5}$$

where the b_i are independent, identically distributed random effects with common $\mathcal{N}(\boldsymbol{0}, \sigma^2\Sigma)$ distribution and

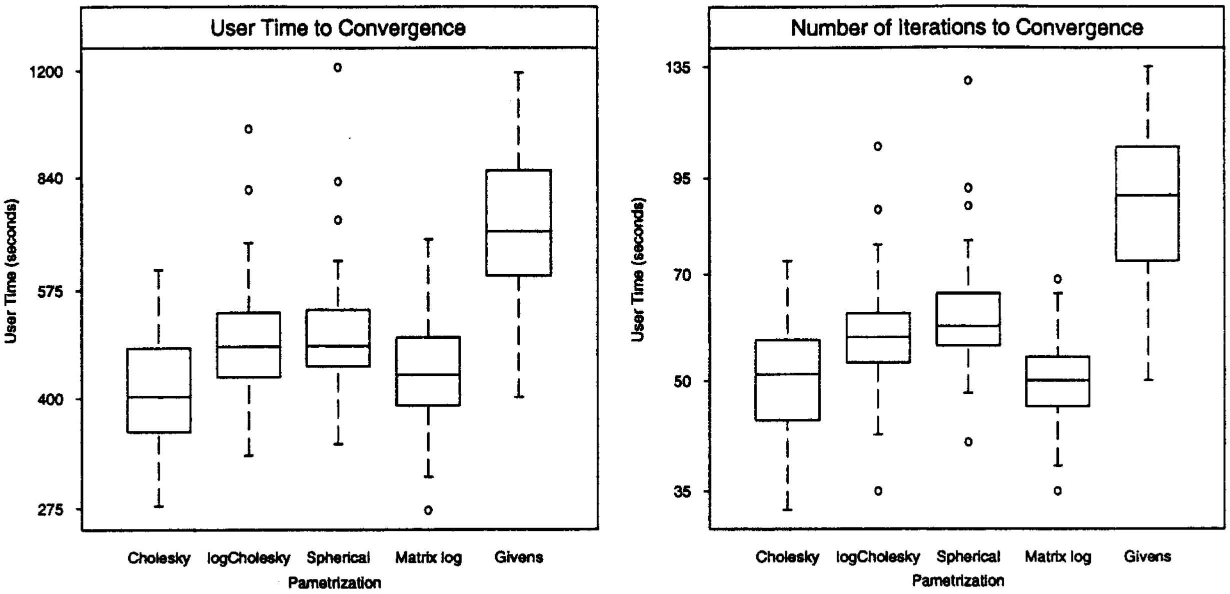

Fig. 3. *Box-plots of user time and number of iterations to convergence for 50 random samples of model (5) with Σ of dimension 6*

the ε_i are independent and identically distributed error terms with common distribution $\mathcal{N}_{n_i}(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$, independent of the $\boldsymbol{b}_i$, with n_i representing the number of observations on the ith cluster. Lindstrom and Bates (1988) have shown that the log-likelihood in (5) can be profiled to produce a function of $\boldsymbol{\Sigma}$ alone. In the simulation, we used $\boldsymbol{\Sigma}$ matrices of dimensions 3 and 6. These were defined such that the non-zero elements of the ith column of the corresponding Cholesky factor were equal to the integers between 1 and i. For $n = 3$ we have $\boldsymbol{\Sigma} = \boldsymbol{A}$, as given in (2). For $n = 3$ we used $M = 10$, $n_i = 15$, $i = 1, \ldots, 10$, $\sigma^2 = 1$, and $\boldsymbol{\beta} = (10, 1, 2)^{\mathrm{T}}$, while for $n = 6$ we used $M = 50$, $n_i = 25$, $i = 1, \ldots, 50$, $\sigma^2 = 1$, and $\boldsymbol{\beta} = (10, 1, 2, 3, 4, 5)^{\mathrm{T}}$. In both cases, the elements of the first column of $\boldsymbol{X}$ were set equal to 1 and the remaining elements were independently generated according to a U(1, 20) distribution. A total of 300 and 50 samples were generated respectively for $n = 3$ and $n = 6$, and the number of iterations and the user time to calculate the maximum likelihood estimate of $\boldsymbol{\Sigma}$ for each parametrization recorded.

Figures 2 and 3 present box-plots of the number of iterations and of the user times for the various parametrizations. The Cholesky, the log-Cholesky, the spherical, and the matrix logarithm parametrizations had similar performances for $n = 3$, considerably better than the Givens parametrization. For $n = 6$ the Cholesky and the matrix logarithm parametrizations gave the best performances, followed by the log-Cholesky and spherical parametrizations, all considerably better than the Givens parametrization. Because $\boldsymbol{\Sigma}$ is relatively small in these examples, the numerical complexity of the different parametrizations did not play a major role in their performances. It is interesting to note that even though the matrix logarithm is one of the least efficient parametrizations in terms of numerical complexity, it had the best performance in terms of number of iterations and user time to obtain the maximum likelihood estimate of $\boldsymbol{\Sigma}$, suggesting that this parametrization is most numerically stable.

Another important aspect in which the parametrizations should be compared is their behaviour as $\boldsymbol{\Sigma}$ approaches singularity. All parametrizations described in Section 2 require $\boldsymbol{\Sigma}$ to be positive definite, though the Givens parametrization can be modified to handle general symmetric matrices. It is usually an important statistical issue to test whether $\boldsymbol{\Sigma}$ is of less than full rank, in which case the dimension of the parameter space can be reduced.

As $\boldsymbol{\Sigma}$ approaches singularity its determinant goes to zero and so at least one of the diagonal elements of its Cholesky factor goes to zero too. The Cholesky parametrization would then become numerically unstable, since equivalent solutions would get closer together in the estimation space. At least one element of $\boldsymbol{\theta}$ in the log-Cholesky parametrization would go to $-\infty$ (the logarithm of the diagonal element of $\boldsymbol{L}$ that goes to zero). In the spherical parametrization we would also have at least one element of $\boldsymbol{\theta}$ going in absolute value to ∞: if the first diagonal element of $\boldsymbol{L}$ goes to zero, $\theta_1 \to -\infty$; otherwise at least one angle of the spherical coordinates of the column of $\boldsymbol{L}$ whose diagonal element approaches 0 would either approach 0 or π, in which cases the corresponding element of $\boldsymbol{\theta}$ would go respectively to $-\infty$ or ∞.

Singularity of $\boldsymbol{\Sigma}$ implies that at least one of its eigenvalues is zero. The Givens parametrization would then have at least the first element of $\boldsymbol{\theta}$ going to $-\infty$. To understand what happens with the matrix logarithm parametrization when $\boldsymbol{\Sigma}$ approaches singularity we note that letting $(\lambda_1, \boldsymbol{u}_1), \ldots, (\lambda_n, \boldsymbol{u}_n)$ represent the eigenvalue–eigenvector pairs corresponding to $\boldsymbol{\Sigma}$ we can write $\boldsymbol{\Sigma} = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^{\mathrm{T}}$. As $\lambda_1 \to 0$ all entries of $\log(\boldsymbol{\Sigma})$ corresponding to non-zero elements of $\boldsymbol{u}_1 \boldsymbol{u}_1^{\mathrm{T}}$ would converge in absolute value to ∞. Hence in the matrix logarithm parametrization we could have all elements of $\boldsymbol{\theta}$ going either to $-\infty$ of ∞ as $\boldsymbol{\Sigma}$ approached singularity.

Finally we consider the statistical interpretability of the parametrizations of $\boldsymbol{\Sigma}$. The least interpretable parametrization is the matrix logarithm—none of its elements can be directly related to the individual variances, covariances, or eigenvalues of $\boldsymbol{\Sigma}$. The Cholesky and log-Cholesky parametrizations have the first component directly related to the variance of X_1, the first underlying random variable in $\boldsymbol{\Sigma}$. By permuting the order of the random variables in the definition of $\boldsymbol{\Sigma}$, one can derive measures of variability and confidence intervals for all the variances in $\boldsymbol{\Sigma}$, from corresponding quantities obtained for the parameters in the Cholesky or log-Cholesky parametrizations. The Givens parametrization is the only one considered here that uses the eigenvalues of $\boldsymbol{\Sigma}$ directly in the definition of $\boldsymbol{\theta}$. It is a very useful parametrization for identifying ill-conditioning of $\boldsymbol{\Sigma}$. None of its parameters, though, can be directly related to the variances and covariances in $\boldsymbol{\Sigma}$. Finally, the spherical parametrization is the one that gives the largest number of interpretable parameters of all parametrizations considered here. Measures of variability and confidence intervals for all the variances in $\boldsymbol{\Sigma}$ and the correlations with X_1 can be obtained from the corresponding quantities calculated for $\boldsymbol{\theta}$. By permuting the order of the underlying random variables in the definition of $\boldsymbol{\Sigma}$, one can in fact derive measures of variability and confidence intervals for all the variances and correlations in $\boldsymbol{\Sigma}$.

4. Conclusions

The parametrizations described in Section 2 allow the estimation of variance–covariance matrices using unconstrained optimization. This has numerical and statistical advantages over constrained optimization, since the latter is usually a much harder numerical problem.

Furthermore, unconstrained estimates tend to have better inferential properties.

Of the five parametrizations considered here, the spherical parametrization presents the best combination of performance and statistical interpretability of individual parameters. The Cholesky and log-Cholesky parametrizations have comparable performances, similar to the spherical parametrization, but lack direct parameter interpretability. The Givens parametrization is considerably less efficient than these parametrizations, but has the feature of being directly based on the eigenvalues of the variance–covariance matrix. This can be used, for example, to identify non-random linear combinations of the underlying random variables. The matrix logarithm parametrization is very inefficient as the dimension of the variance–covariance matrix increases, but seems to be the most stable parametrization. It also lacks direct interpretability of its parameters.

Different parametrizations can be used at different stages of the data analysis. The matrix logarithm parametrization seems to be the most efficient for the optimization step, at least for moderately large Σ. The spherical parametrizations is probably the best one to derive measures of variability and confidence intervals for the elements of Σ, while the Givens parametrization is the most convenient to investigate rank deficiency of Σ.

Only unstructured variance–covariance matrices were considered here, but in many situations that involve the optimization of an objective function, structured matrices are used instead (Jennrich and Schluchter, 1986). It is therefore important to derive parametrizations for structured variance–covariance matrices that allow unconstrained estimation of the associated parameters.

The asymptotic properties of the different parametrizations considered here have not yet been studied and certainly constitute an interesting research topic. It may be that some of the parametrizations give faster rates of convergence to normality than others and this could be used as a criterion for choosing among them.

Acknowledgements

This research was partially supported by the National Science Foundation through grant number DMS-9309101 and by the Coordenação para Aperfeiçoamento de Pessoal de Nível Superior, Brazil. We are grateful to Bruce E. Ankenman for suggesting the Givens parametrization and to two anonymous referees for helpful comments and suggestions.

References

Anderson, T. W., Olkin, I. and Underhill, L. G. (1987) Generation of random orthogonal matrices. *SIAM Journal on Scientific and Statistical Computing*, **8**(4), 625–9.

Bates, D. M. and Watts, D. G. (1988) *Nonlinear Regression Analysis and Its Applications*. Wiley, New York.

Dennis, Jr., J. E. and Schnabel, R. B. (1983) *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*. Prentice-Hall, Englewood Cliffs, NJ.

Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated measures models with structural covariance matrices. *Biometrics*, **42**(4), 805–20.

Jupp, D. L. B. (1978) Approximation to data by splines with free knots. *SIAM Journal of Numerical Analysis*, **15**(2), 328–43.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–74.

Leonard, T. and Hsu, J. S. J. (1993) Bayesian inference for a covariance matrix. *Annals of Statistics*, **21**, 1–25.

Lindstrom, M. J. and Bates, D. M. (1988) Newton–Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, **83**, 1014–22.

Lindstrom, M. J. and Bates, D. M. (1990) Nonlinear mixed effects models for repeated measures data. *Biometrics*, **46**, 673–87.

Pinheiro, J. C. (1994) Topics in Mixed Effects Models. PhD thesis, University of Wisconsin–Madison.

Pinheiro, J. C. and Bates, D. M. (1995) Model building for nonlinear mixed-effects models. Technical Report 91, Department of Biostatistics, University of Wisconsin–Madison.

Rao, C. R. (1973) *Linear Statistical Inference and Its Applications*, 2nd edn. Wiley, New York.

Thisted, R. A. (1988) *Elements of Statistical Computing*. Chapman & Hall, London.