ELSEVIER Linear Algebra and its Applications 281 (1998) 97–103

LINEAR ALGEBRA AND ITS APPLICATIONS

Approximating the inverse of a symmetric positive definite matrix

Gordon Simons [a,*], Yi-Ching Yao [b]

[a] Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA
[b] Institute of Statistical Science, Academia Sinica, Taipei, Taiwan

Received 29 August 1997; received in revised form 26 February 1998; accepted 3 March 1998

Submitted by G.P.H. Styan

Abstract

It is shown for an $n \times n$ symmetric positive definite matrix $T = (t_{i,j})$ with negative off-diagonal elements, positive row sums and satisfying certain bounding conditions that its inverse is well approximated, uniformly to order $1/n^2$, by a matrix $S = (s_{i,j})$, where $s_{i,j} = \delta_{i,j}/t_{i,i} + 1/t_{..}$, $\delta_{i,j}$ being the Kronecker delta function, and $t_{..}$ being the sum of the elements of T. An explicit bound on the approximation error is provided. © 1998 Elsevier Science Inc. All rights reserved.

1. Introduction

We are concerned here with $n \times n$ symmetric matrices $T = (t_{i,j})$ which have negative off-diagonal elements and positive row (and column) sums, i.e.,

$$t_{i,j} = t_{j,i},\ t_{i,j} < 0 \quad \text{for } i \neq j \quad \text{and} \quad \sum_{k=1}^{n} t_{i,k} > 0 \quad \text{for } i, j = 1, \ldots, n.$$

Such matrices must be positive definite and hence fall into the class of M-matrices. (See, e.g., [1] for the definition and properties of M-matrices.)

It is convenient to introduce an array $\{u_{i,j}\}_{i,j=1}^{n}$ of positive numbers defined in terms of T as follows:

* Corresponding author. E-mail: simons@stat.unc.edu.

PII: S0024-3795(98)10038-1

$$u_{i,j} = -t_{i,j} \quad \text{for} \quad i \neq j \quad \text{and} \quad u_{i,i} = \sum_{k=1}^{n} t_{i,k}, \quad i,j = 1,\ldots,n.$$

Then we have

$$u_{i,j} > 0,\ u_{i,j} = u_{j,i}, \quad t_{i,j} = -u_{i,j} \quad \text{for } i \neq j, \quad \text{and}$$
$$t_{i,i} = \sum_{k=1}^{n} u_{i,k}, \quad i,j = 1,\ldots,n. \tag{1}$$

Moreover, it is convenient to introduce the notation

$$m = \min_{i,j} u_{i,j}, \qquad M = \max_{i,j} u_{i,j}, \qquad t_{..} = \sum_{i,j=1}^{n} t_{i,j} = \sum_{k=1}^{n} u_{k,k} > 0, \tag{2}$$

$\|A\| = \max_{i,j} \mid a_{i,j} \mid$ for a general matrix $A = (a_{i,j})$, and the $n \times n$ symmetric positive definite matrix $S = (s_{i,j})$, with

$$s_{i,j} = \frac{\delta_{i,j}}{t_{i,i}} + \frac{1}{t_{..}},$$

where $\delta_{i,j}$ denotes the Kronecker delta function.

Theorem.

$$\|T^{-1} - S\| \leqslant \frac{C(m,M)}{n^2},$$

where

$$C(m,M) = \left(1 + \frac{M}{m}\right)\frac{M}{m^2}.$$

The authors [2] use this theorem while establishing the asymptotic normality of a vector-valued estimator arising in a study of the Bradley–Terry model for paired comparisons. Depending on n, which goes to infinity in the asymptotic limit, we need to consider the inverse T^{-1} of a matrix T satisfying Eq. (1) with m and M being bounded away from 0 and infinity. Since it is impossible to obtain this inverse explicitly, except for a few special cases, we show that the approximate inverse S is a workable substitute, with the attendant errors going to zero at the rate $1/n^2$ as $n \to \infty$.

Computing and estimating the inverse of a matrix has been extensively studied and described in the literature. See [3–5] and references therein. In [4], the characterization of inverses of symmetric tridiagonal and block tridiagonal matrices is discussed, which gives rise to stable algorithms for computing their inverses. [3] and [5] derive, among other things, upper and lower bounds for the elements of the inverse of a symmetric positive definite matrix. In particular, for a symmetric positive definite matrix $A = (a_{i,j})$ of dimension

n, the following bounds on the diagonal elements of A^{-1} are given in [3] and [5]:

$$\frac{1}{\alpha}+\frac{(\alpha-a_{i,i})^2}{\alpha(\alpha a_{i,i}-\sum_{k=1}^n a_{i,k}^2)}\leqslant (A^{-1})_{i,i}\leqslant \frac{1}{\beta}-\frac{(a_{i,i}-\beta)^2}{\beta(\sum_{k=1}^n a_{i,k}^2-\beta a_{i,i})},$$

where $\alpha \geqslant \lambda_n$ and $0<\beta\leqslant\lambda_1$, λ_1 and λ_n being the smallest and largest eigenvalues of A, respectively.

The next section contains the proof of the theorem, and some remarks are given in Section 3.

2. Proof of the theorem

Note that

$$T^{-1}-S=(T^{-1}-S)(I_n-TS)+S(I_n-TS),$$

where I_n is the $n\times n$ identity matrix. Letting $V=I_n-TS$ and $W=SV$, we have

$$T^{-1}-S=(T^{-1}-S)V+W.$$

Thus the task is to show that $\|F\|\leqslant C(m,M)$, where the matrices $F=n^2(T^{-1}-S)$ and $G=n^2W$ satisfy the recursion

$$F=FV+G. \tag{3}$$

By the definitions of S, $V=(v_{i,j})$ and $W=(w_{i,j})$, it follows from Eqs. (1) and (2) that

$$\begin{aligned} v_{i,j}&=\delta_{i,j}-\sum_{k=1}^n t_{i,k}s_{k,j}\\ &=\delta_{i,j}-\sum_{k=1}^n t_{i,k}\left(\frac{\delta_{k,j}}{t_{j,j}}+\frac{1}{t_{..}}\right)\\ &=\delta_{i,j}-\frac{t_{i,j}}{t_{j,j}}-\frac{u_{i,i}}{t_{..}}\\ &=(1-\delta_{i,j})\,\frac{u_{i,j}}{t_{j,j}}-\frac{u_{i,i}}{t_{..}} \end{aligned}\tag{4}$$

and

$$\begin{aligned} w_{i,j}&=\sum_{k=1}^n s_{i,k}v_{k,j}=\sum_{k=1}^n\left(\frac{\delta_{i,k}}{t_{i,i}}+\frac{1}{t_{..}}\right)\left((1-\delta_{k,j})\frac{u_{k,j}}{t_{j,j}}-\frac{u_{k,k}}{t_{..}}\right)\\ &=\sum_{k=1}^n\frac{\delta_{i,k}}{t_{i,i}}\left((1-\delta_{k,j})\frac{u_{k,j}}{t_{j,j}}-\frac{u_{k,k}}{t_{..}}\right)+\sum_{k=1}^n\frac{1}{t_{..}}\left((1-\delta_{k,j})\frac{u_{k,j}}{t_{j,j}}-\frac{u_{k,k}}{t_{..}}\right)\\ &=\frac{(1-\delta_{i,j})u_{i,j}}{t_{i,i}t_{j,j}}-\frac{u_{i,i}}{t_{i,i}t_{..}}-\frac{u_{j,j}}{t_{j,j}t_{..}}. \end{aligned}\tag{5}$$

Again by Eqs. (1) and (2), we have

$$0 < \frac{u_{i,j}}{t_{i,i}t_{j,j}} \leqslant \frac{M}{m^2n^2}, \qquad 0 < \frac{u_{i,i}}{t_{i,i}t_{..}} \leqslant \frac{M}{m^2n^2},$$

so that

$$|w_{i,j}| \leqslant \frac{a}{n^2} \text{ and } |w_{i,j} - w_{i,k}| \leqslant \frac{a}{n^2} \quad \text{for } i,j,k = 1,\ldots,n,$$

where $a = 2M/m^2$. Equivalently, in terms of the elements of $G = (g_{i,j})$:

$$|g_{i,j}| \leqslant a \text{ and } |g_{i,j} - g_{i,k}| \leqslant a, \quad i,j,k = 1,\ldots,n. \tag{6}$$

We now turn our attention to Eq. (3), expressed in terms of the matrix elements $f_{i,j}$ and $g_{i,j}$ in F and G, respectively, and the formula for $v_{i,j}$ in Eq. (4):

$$f_{i,j} = \sum_{k=1}^{n} f_{i,k}(1 - \delta_{k,j})\frac{u_{k,j}}{t_{j,j}} - \sum_{k=1}^{n} f_{i,k}\frac{u_{k,k}}{t_{..}} + g_{i,j}, \quad i,j = 1,\ldots,n. \tag{7}$$

The task is to show $|f_{i,j}| \leqslant C(m,M)$ for all i and j.

Two things are readily apparent in Eq. (7). To begin with, apart from the factor $(1 - \delta_{k,j})$ in the first sum, which equals one except when $k = j$, the first and second sums are weighted averages of $f_{i,k}$, $k = 1,\ldots,n$; the positive weights $u_{k,j}/t_{j,j}$ and $u_{k,k}/t_{..}$ each add to unity in the index k. Secondly, the index i plays no essential role in the relationship; it can be viewed as fixed. If we take i to be fixed and notationally suppress it in Eq. (7), then Eq. (7) assumes the form of n linear equations in the n unknowns $f_1,\ldots,f_n$:

$$f_j = \sum_{k=1}^{n} f_k(1 - \delta_{k,j})\frac{u_{k,j}}{t_{j,j}} - \sum_{k=1}^{n} f_k\frac{u_{k,k}}{t_{..}} + g_j, \quad j = 1,\ldots,n. \tag{8}$$

Instead of solving these equations, we will show that under the bounding conditions

$$|g_j| \leqslant a, \quad |g_j - g_k| \leqslant a, \; j,k = 1,\ldots,n,$$

(see Eq. (6)) any solution of Eq. (8) must satisfy the inequalities

$$|f_j| \leqslant \frac{1}{2}\left(1 + \frac{M}{m}\right)a, \quad j = 1,\ldots,n, \tag{9}$$

so that $|f_j| \leqslant C(m,M)$, $j = 1,\ldots,n$, thereby completing the proof.

Let α and β be such that $f_\alpha = \max_{1 \leqslant k \leqslant n} f_k$ and $f_\beta = \min_{1 \leqslant k \leqslant n} f_k$. With no loss of generality, assume $f_\alpha \geqslant |f_\beta|$. (Otherwise, we may reverse the signs of the f_k's and proceed analogously.) There are two cases to consider:

Case I: $f_\beta \geqslant 0$. Then

$$\begin{aligned}
f_\alpha &= \sum_{k=1}^{n} f_k(1-\delta_{k,\alpha})\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \sum_{k=1}^{n} f_k\frac{u_{k,k}}{t_{..}} + g_\alpha \\
&\leqslant \sum_{k=1}^{n} f_k\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \sum_{k=1}^{n} f_k\frac{u_{k,k}}{t_{..}} + g_\alpha \\
&= \sum_{k=1}^{n} f_k\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,k}}{t_{..}}\right) + g_\alpha \\
&\leqslant f_\alpha \sum_{k\in A}\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,k}}{t_{..}}\right) + g_\alpha,
\end{aligned}$$

where $A = \{k : u_{k,\alpha}/t_{\alpha,\alpha} > u_{k,k}/t_{..}\}$. Let ρ denote the cardinality of A, and observe that

$$\sum_{k\in A}\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,k}}{t_{..}}\right) \leqslant \frac{M\rho}{M\rho + m(n-\rho)} - \frac{m\rho}{m\rho + M(n-\rho)} \leqslant \frac{M-m}{M+m}, \tag{10}$$

the first inequality being an immediate consequence of the constraints $m \leqslant u_{i,j} \leqslant M$ (see Eq. (2)) and the sum formulas in Eqs. (1) and (2), the second inequality taking into account that the middle expression in Eq. (10) is a concave function of ρ (when viewed as a continuous variable between 0 and n), with its maximum occurring at $\rho = n/2$. Thus,

$$f_\alpha \leqslant f_\alpha \sum_{k\in A}\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,k}}{t_{..}}\right) + g_\alpha \leqslant f_\alpha\frac{M-m}{M+m} + g_\alpha \leqslant f_\alpha\frac{M-m}{M+m} + a,$$

so that

$$f_\alpha \leqslant \frac{1}{2}\left(1 + \frac{M}{m}\right) a = C(m, M),$$

thereby establishing Eq. (9) and completing the proof.

Case II: $f_\beta < 0$. Let $h_k = f_k - f_\beta \geqslant 0$, $k = 1, \ldots, n$. Then

$$\begin{aligned}
h_\alpha &= f_\alpha - f_\beta \\
&\leqslant \sum_{k=1}^{n} f_k\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \sum_{k=1}^{n} f_k\frac{u_{k,\beta}}{t_{\beta,\beta}} + g_\alpha - g_\beta \\
&= \sum_{k=1}^{n} h_k\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \sum_{k=1}^{n} h_k\frac{u_{k,\beta}}{t_{\beta,\beta}} + g_\alpha - g_\beta \\
&= \sum_{k=1}^{n} h_k\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,\beta}}{t_{\beta,\beta}}\right) + g_\alpha - g_\beta \\
&\leqslant h_\alpha \sum_{k\in A}\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}} - \frac{u_{k,\beta}}{t_{\beta,\beta}}\right) + g_\alpha - g_\beta,
\end{aligned}$$

where $A = \{k : u_{k,\alpha}/t_{\alpha,\alpha} > u_{k,\beta}/t_{\beta,\beta}\}$. The argument from this point proceeds analogously to that for Case I. Letting ρ denote the cardinality of A, one obtains

$$\sum_{k\in A}\left(\frac{u_{k,\alpha}}{t_{\alpha,\alpha}}-\frac{u_{k,\beta}}{t_{\beta,\beta}}\right)\leqslant\frac{M\rho}{M\rho+m(n-\rho)}-\frac{m\rho}{m\rho+M(n-\rho)}\leqslant\frac{M-m}{M+m},$$

which leads to

$$h_\alpha\leqslant h_\alpha\frac{M-m}{M+m}+g_\alpha-g_\beta\leqslant h_\alpha\frac{M-m}{M+m}+a,$$

so that

$$f_\alpha\leqslant h_\alpha\leqslant\frac{1}{2}\left(1+\frac{M}{m}\right)a,$$

thereby establishing Eq. (9) and completing the proof. □

3. Remarks

While our proof of the theorem is somewhat long, we do not see how to simplify it by using any of the well-known properties of M-matrices.

The bound $C(m,M)/n^2$ on the approximation error is a product of two factors, one depending on m and M, the other on n. For large n, with m and M held bounded away from 0 and infinity, the elements of S (and hence of T^{-1}) are all of order $1/n$, and the errors (i.e., the elements of $T^{-1}-S$) are uniformly $O(1/n^2)$ as $n\to\infty$. This fact is crucially used in Ref. [2].

A particular case of the matrix T, described below, shows that the factor $1/n^2$ is best possible in the sense that any bound of the for $\tilde{C}(m,M)/\gamma(n)$ requires $\gamma(n)=O(n^2)$ as $n\to\infty$; no faster growth rate than n^2 is allowed. On the other hand, it is natural to ask whether the factor $C(m,M)$ is best possible. To clarify the issue, for given integer n and given m and M, $0<m\leqslant M<\infty$, let $Q_n(m,M)$ denote the set of $n\times n$ symmetric positive definite matrices satisfying (1) with $m\leqslant u_{i,j}\leqslant M$, $i,j=1,\ldots,n$, and define

$$C_o(m,M)=\sup\{n^2\|T^{-1}-S\| : T\in Q_n(m,M),\ n=1,2,\ldots\}.$$

It follows from the theorem that $C_o(m,M)\leqslant C(m,M)=(1+M/m)M/m^2$. But for the special matrix T satisfying Eq. (1) with $u_{1,1}=M$ and $u_{i,j}=m$ for all other (i,j), we find that

$$
(T^{-1})_{i,j} = \begin{cases} \frac{2}{2M+(n-1)m} & \text{for } i = j = 1, \\ \frac{1}{2M+(n-1)m} & \text{for } i = 1, j \neq 1 \text{ or } i \neq 1, j = 1, \\ \frac{3M+(2n-1)m}{(n+1)m(2M+(n-1)m)} & \text{for } i = j \neq 1, \\ \frac{M+nm}{(n+1)m(2M+(n-1)m)} & \text{for } 1 \neq i \neq j \neq 1. \end{cases}
$$

So

$$
(T^{-1} - S)_{1,1} = \frac{-2M}{(M + (n-1)m)(2M + (n-1)m)},
$$

from which it follows that $C_o(m, M) \geqslant 2M/m^2$. The same matrix T justifies the constraint on $\gamma(n)$ described above.

The gap between $2M/m^2$ and $(1 + M/m)M/m^2$ suggests that there might be room for improvement in our bound. Indeed, by computer, we have numerically inverted a very large number of matrices of various dimensions (some as large as 300×300) and found that the inequality $n^2\|T^{-1} - S\| \leqslant 2M/m^2$ holds in all cases. It would therefore be interesting to see whether $C_o(m, M) = 2M/m^2$.

We finish with one final observation. Surprisingly, it is possible to evaluate the second sum in Eq. (7) explicitly:

$$
\sum_{k=1}^{n} f_{i,k} \frac{u_{k,k}}{t_{..}} = -n^2 \frac{u_{i,i}}{t_{i,i} t_{..}},
$$

which is identical to, and permits a cancellation with, one of the three terms defining $g_{i,j}$ (cf., Eq. (5)). To obtain this, one multiplies both sides of Eq. (7) by $t_{j,j}$, adds over j ($j = 1, \ldots, n$), and carries out the suggested algebra. While we have not found much use for this identity, it does show that f_β, appearing in the proof of the theorem, is strictly negative. Since, as it turns out, f_α can be positive or negative, neither case described in the proof is superfluous.

References

[1] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, Cambridge University Press, New York, 1991.

[2] G. Simons, Y.-C. Yao, A large sample study of the Bradley-Terry model for paired comparisons, in preparation.

[3] G.H. Golub, Z. Strakoš, Estimates in quadratic formulas, Numer. Algorithms 8 (1994) 241–268.

[4] G. Meurant, A review of the inverse of symmetric tridiagonal and block tridiagonal matrices, SIAM J. Matrix Anal. Appl. 13 (1992) 707–728.

[5] P.D. Robinson, A.J. Wathen, Variational bounds on the entries of the inverse of a matrix, IMA J. Numer. Anal. 12 (1992) 463–486.