Linear estimation in models based on a graph

R.B. Bapat [*]

Indian Statistical Institute, New Delhi 110 016, India

Received 2 September 1998; accepted 15 April 1999

Submitted by B. Datta

Dedicated to Hans Schneider

Abstract

Two natural linear models associated with a graph are considered. The Gauss–Markov theorem is used in one of the models to derive a combinatorial formula for the Moore–Penrose inverse of the incidence matrix of a tree. An inequality involving the Moore–Penrose inverse of the Laplacian matrix of a graph and its distance matrix is obtained. The case of equality is discussed. Again the main tool used in the proof is the theory of linear estimation.

AMS classification: Primary: 15A09; Secondary: 05C05

Keywords: Linear model; Moore–Penrose inverse; Tree; Incidence matrix; Laplacian matrix; Distance

1. Preliminaries

A *graph* $G = (V, E)$ consists of a finite set of vertices, V, and a set of edges, E. Each edge is a pair of distinct vertices. We consider graphs which have no loops or multiple edges. For basic graph-theoretic notions we refer to [4].

A *directed graph* is a graph in which each edge has been assigned an orientation. Let G be a directed graph with $V = \{1, \dots, n\}$, $E = \{e_1, \dots, e_m\}$. The incidence matrix of G, denoted by Q, is the $n \times m$ matrix defined as follows.

The (i,j)-entry of Q is 0 if vertex i and edge e_j are not incident and otherwise it is 1 or -1 according as e_j originates or terminates at i, respectively.

For a directed graph G, the matrix $L = QQ^{\mathrm{T}}$ is called the *Laplacian matrix* (or the *Kirchhoff matrix*) of G. Note that the Laplacian matrix does not depend on the orientation of G and hence is essentially defined for an undirected graph. The matrix $K = Q^{\mathrm{T}}Q$ has been called the *edge version* of the Laplacian matrix.

If $G = (V, E)$ is a connected graph (directed or otherwise) and if $i, j \in V$, then the distance between i, j, denoted by d_{ij}, is defined as the length (i.e., the number of edges) in a shortest path between i and j. (When we talk of a path or a cycle in G, we mean a path or a cycle in the underlying undirected graph. These notions of path and cycle differ from the standard digraph notions, where, for example, in a path from i to j, the arcs must be oriented from i to j.) The distance matrix $D = [d_{ij}]$ has been considered in the literature as well. In particular, when the graph is a tree, the distance matrix is closely related to the Laplacian and its edge version. For several properties of the Laplacian matrix, the edge version of the Laplacian and the distance matrix we refer to the papers by Merris [12,13], and the references contained therein.

We now recall some basic aspects of the theory of linear models. Suppose $Y_1, \ldots, Y_n$ are random variables such that the expectation of each Y_i is a linear combination of certain parameters $\beta_1, \ldots, \beta_p$. We can express this information as a linear model $E(Y) = X\beta$, where $E(Y)$ denotes the expectation of the vector $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, X is an $n \times p$ matrix of (known) coefficients and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$. We also assume that $Y_1, \ldots, Y_n$ are uncorrelated with a common unknown variance σ^2. Thus the dispersion matrix of Y, denoted by $D(Y)$, is given by $D(Y) = \sigma^2 I_n$, where I_n is the $n \times n$ identity matrix. For a discussion of linear models, including concepts such as estimability and best linear unbiased estimate (BLUE) see [15,1].

If A is an $n \times m$ matrix, then an $m \times n$ matrix B is called a generalized inverse of A if $ABA = A$. The *Moore–Penrose inverse* of A, denoted by A^+, is an $m \times n$ matrix satisfying the equations $ABA = A$, $BAB = B$, $(AB)^{\mathrm{T}} = AB$ and $(BA)^{\mathrm{T}} = BA$. It is well-known that any real or complex matrix admits a unique Moore–Penrose inverse. We refer to [3,5] for basic properties of the Moore–Penrose inverse. For some recent results concerning the Moore–Penrose inverse of a Laplacian, see [6–8].

2. The Moore–Penrose inverse of the incidence matrix of a tree

A graph-theoretic description of the Moore–Penrose inverse of the incidence matrix of a directed tree was recently given in [2]. The formula was used to obtain an expression for the inverse of the edge version of the Laplacian, K, derived earlier by Moon [14] and by Merris [12].

In this section we consider a linear model where the coefficient matrix is the incidence matrix of a tree. The standard Gauss–Markov theorem is then used to derive a graph-theoretic description of the Moore–Penrose inverse of the incidence matrix Q.

Our underlying graph is assumed to be a directed tree. Let $T=(V,E)$ be a directed tree with $V=\{1,\ldots,n\}$ and $E=\{e_1,\ldots,e_m\}$. Note that $m=n-1$. Let Q be the incidence matrix of T. It is well-known (see, for example [4]) that the rank of Q is $m=n-1$. Thus the linear model $E(Y)=Q\beta,\ D(Y)=\sigma^2 I_n$ is a full rank model. In particular, each β_i is *estimable*, i.e., there exists ℓ such that $E(\ell^{\mathrm{T}}Y)=\beta_i$.

If $e_i\in E$ then observe that the graph $T\setminus\{e_i\}$ has two components, both being trees. This observation is relevant in the statement of the next result.

Theorem 1. *Let $T=(V,E)$ be a directed tree with $V=\{1,\ldots,n\}$ and $E=\{e_1,\ldots,e_{n-1}\}$. Let Q be the incidence matrix of T and let $Q^+=[q^+_{ij}]$ be the Moore–Penrose inverse of Q. Then $n|q^+_{ij}|$ equals the number of vertices in the component of $T\setminus e_i$ not containing j. Furthermore, q^+_{ij} is positive or negative according as e_i is directed away from j or towards j, respectively.*

Proof. Consider the linear model $E(Y)=Q\beta,\ D(Y)=\sigma^2 I_n$. As observed earlier, this is a full rank model and hence each β_i is estimable. Let $\hat{\beta}_i$ denote the BLUE of $\beta_i,\ i=1,\ldots,n-1$, and let $\hat{\beta}=(\hat{\beta}_1,\ldots,\hat{\beta}_{n-1})^{\mathrm{T}}$. By the Gauss–Markov theorem, $\hat{\beta}=(Q^{\mathrm{T}}Q)^{-1}Q^{\mathrm{T}}Y$, and since $Q^+=(Q^{\mathrm{T}}Q)^{-1}Q^{\mathrm{T}}$, we have $\hat{\beta}=Q^+Y$. Thus if $c_1Y_1+\cdots+c_nY_n$ is the BLUE of β_i, then $[c_1,\ldots,c_n]$ gives the ith row of $Q^+,\ i=1,\ldots,n-1$.

We now find the BLUE of β_i. If $c^{\mathrm{T}}Y$ is unbiased for β_i, then $E(c^{\mathrm{T}}Y)=c^{\mathrm{T}}Q\beta=\beta_i$ for any β, or equivalently,

$$c^{\mathrm{T}}Q\beta=(0,\ldots,0,1,0,\ldots,0)\beta,$$

where the 1 occurs at the ith place. Since β is arbitrary, we conclude that

$$c^{\mathrm{T}}Q=(0,\ldots,0,1,0,\ldots,0).$$

Thus we have

$$c_k-c_\ell=\begin{cases}1 & \text{if } (k,\ell)=e_i,\\ 0 & \text{if } (k,\ell)\in E,\ (k,\ell)\neq e_i.\end{cases}\tag{1}$$

Let us suppose that edge e_i joins vertices p,q and that it is directed from p to q. Let $T_1,\ T_2$ be the components of $T\setminus\{e_i\}$ and let $V_1,\ V_2$ be the corresponding vertex sets, respectively. We assume, without loss of generality, that $p\in V_1,\ q\in V_2$.

From (1) we conclude that there exists α such that $c_p=1+\alpha,\ c_q=\alpha$ and

$$c_j=\begin{cases}1+\alpha & \text{if } j\in V_1,\\ \alpha & \text{if } j\in V_2.\end{cases}\tag{2}$$

In order to find the BLUE of β_i, we must minimize

$$c_1^2 + \cdots + c_n^2 = (1+\alpha)^2|V_1| + \alpha^2|V_2|.$$

Setting the derivative with respect to α, equal to zero, we get

$$2(1+\alpha)|V_1| + 2\alpha|V_2| = 0$$

and hence

$$\alpha = -\frac{|V_1|}{|V_1| + |V_2|} = -\frac{|V_1|}{n}.$$

Substituting in (2) we find the linear estimator $c^{\mathrm{T}}Y$ which is the BLUE of β_i. Thus if $j \in V_1$, then $c_j = |V_2|/n$ while if $j \in V_2$, then $c_j = -|V_1|/n$. This establishes the result. □

We remark that Theorem 1 as well as the results in Section 3 continue to hold for weighted graphs (i.e., graphs in which each edge is assigned a positive weight). This only requires obvious modifications in the statements and the proofs. We deal with the unweighted case for convenience.

3. The main result

In this section we consider graphs which are not necessarily trees. Let $G = (V, E)$ be a directed graph with $V = \{1, \ldots, n\}$, $E = \{e_1, \ldots, e_m\}$. Suppose $\mathscr{P}$ is an (i, j)-path. The incidence vector u of $\mathscr{P}$ is an $m \times 1$ vector defined as follows. The kth entry of u is zero if e_k is not in $\mathscr{P}$. Otherwise it is 1 or -1 according as e_k is directed towards j or away from j, respectively. The incidence vector of a cycle is defined similarly. However, in the case of a cycle we must fix an orientation for the cycle before defining its incidence vector. The choice of the orientation is arbitrary as long as it is kept fixed throughout.

Let G be a directed graph with $V = \{1, \ldots, n\}$, $E = \{e_1, \ldots, e_m\}$ and suppose G has p connected components. Let Q be the incidence matrix of G and consider the linear model

$$E(Y) = Q^{\mathrm{T}}\beta, \quad D(Y) = \sigma^2 I_m.$$

Recall that a function $c^{\mathrm{T}}Y$ is called an error function if $E(c^{\mathrm{T}}Y) = 0$ for any β. Thus $c^{\mathrm{T}}Y$ is an error function if and only if $Qc = 0$. The null space of Q has dimension $m - n + p$. Furthermore, there exists a set of cycles, $\mathscr{C}_1, \ldots, \mathscr{C}_{m-n+p}$, called fundamental cycles, whose incidence vectors form a basis for the null space of Q (see [4, Ch. 12]).

We now turn to estimable functions in this model. A function $\ell^{\mathrm{T}}\beta$ is estimable if and only if ℓ^{T} is in the row space of Q^{T}, or equivalently, ℓ is in the column space of Q. We will often use the fact that $u^{\mathrm{T}}Y$ is the BLUE of the estimable function $\ell^{\mathrm{T}}\beta$ if and only if $E(u^{\mathrm{T}}Y)=\ell^{\mathrm{T}}\beta$ and $\operatorname{cov}(u^{\mathrm{T}}Y, c^{\mathrm{T}}Y)=0$ for any error function $c^{\mathrm{T}}Y$.

The following is the main result of the paper. It is motivated by a result due to Krafft and Schaefer [11], see the discussion given in the end of the paper. We denote the cardinality of the set S by $|S|$. If S denotes a path, a cycle etc., then $|S|$ means the number of edges in S.

Theorem 2. *Let G be a graph with $V=\{1,\dots,n\}$, $E=\{e_1,\dots,e_m\}$. Let L be the Laplacian of G and let $M=L^+$. Let $i,j\in V$ be fixed, $i\neq j$ and let $\mathscr{P}$ be an (i,j)-path of length $\lambda_{ij}>0$. Suppose $\mathscr{C}$ is a cycle in G with $\lambda>0$ edges which satisfies $|\mathscr{P}\cap\mathscr{C}|=t_{ij}$. Then*

$$m_{ii}+m_{jj}-2m_{ij}\leqslant\lambda_{ij}-\frac{t_{ij}^2}{\lambda}. \tag{3}$$

Furthermore, equality holds in (3) if and only if any (i,j)-path is contained in $\mathscr{P}\cup\mathscr{C}$.

Proof. Assign an orientation to G and let Q be the incidence matrix. As before, consider the linear model $E(Y)=Q^{\mathrm{T}}\beta,\ D(Y)=\sigma^2 I_m$. Let $u,\ v$ be the incidence vectors of $\mathscr{P}$, $\mathscr{C}$, respectively. Then for any real α,

$$E(u^{\mathrm{T}}Y+\alpha v^{\mathrm{T}}Y)=E(u^{\mathrm{T}}Y)=\beta_i-\beta_j,$$

and thus $u^{\mathrm{T}}Y+\alpha v^{\mathrm{T}}Y$ is unbiased for $\beta_i-\beta_j$. Therefore

$$\sigma^2(m_{ii}+m_{jj}-2m_{ij})\leqslant\operatorname{var}(u^{\mathrm{T}}Y+\alpha v^{\mathrm{T}}Y) \tag{4}$$

for any real α. The value α_0 of α which minimizes the right-hand side of (4) is seen to be

$$\alpha_0=-\frac{\operatorname{cov}(u^{\mathrm{T}}Y,v^{\mathrm{T}}Y)}{\operatorname{var}(v^{\mathrm{T}}Y)}=-\frac{u^{\mathrm{T}}v}{v^{\mathrm{T}}v}.$$

Setting $\alpha=\alpha_0$ in (4) we get

$$\sigma^2(m_{ii}+m_{jj}-2m_{ij})\leqslant\operatorname{var}(u^{\mathrm{T}}Y)-\frac{(u^{\mathrm{T}}v)^2}{v^{\mathrm{T}}v}. \tag{5}$$

Since $\operatorname{var}(u^{\mathrm{T}}Y)=\sigma^2\lambda_{ij},\ (u^{\mathrm{T}}v)^2=\sigma^2 t_{ij}^2$ and $v^{\mathrm{T}}v=\sigma^2\lambda$, (3) follows from (5).

We now turn to the case of equality. First suppose (3) is strict. Then $(u+\alpha_0 v)^{\mathrm{T}}Y$ is not the BLUE of $\beta_i-\beta_j$. Let the BLUE of $\beta_i-\beta_j$ be $(u+\alpha_0 v+c)^{\mathrm{T}}Y$, where $c^{\mathrm{T}}Y$ is an error function. Then c is in the span of the incidence vectors of fundamental cycles. Suppose c is a linear combination of the incidence vectors of the cycles $\mathscr{C}_1,\dots,\mathscr{C}_k$, each of these appearing with a nonzero coefficient in the linear combination. If none of the cycles $\mathscr{C}_1, \ldots, \mathscr{C}_k$ meet $\mathscr{P} \cup \mathscr{C}$, then clearly,

$$\operatorname{var}((u + \alpha_0 v + c)^{\mathrm{T}} Y) = \operatorname{var}((u + \alpha_0 v)^{\mathrm{T}} Y) + \operatorname{var}(c^{\mathrm{T}} Y) > \operatorname{var}((u + \alpha_0 v)^{\mathrm{T}} Y),$$

contradicting the fact that $(u + \alpha_0 v + c)^{\mathrm{T}} Y$ is BLUE. Thus there must be a cycle $\mathscr{C}_i$ which meets $\mathscr{P} \cup \mathscr{C}$. Also, we may assume that $\mathscr{C}_i \neq \mathscr{C}$. For, if $\mathscr{C}_i = \mathscr{C}$ and if it is the only cycle among $\mathscr{C}_1, \ldots, \mathscr{C}_k$ that meets $\mathscr{P} \cup \mathscr{C}$, then we get a contradiction in view of the choice of α_0. It follows that there is an (i, j)-path not contained in $\mathscr{P} \cup \mathscr{C}$.

Conversely, suppose there is an (i, j)-path, say $\mathscr{P}'$, not contained in $\mathscr{P} \cup \mathscr{C}$. Then there exists a cycle $\mathscr{C}'$ contained in $\mathscr{P} \cup \mathscr{P}'$ such that $\mathscr{C} \cap \mathscr{C}' \subset \mathscr{C} \cap \mathscr{P}$. Let w be the incidence vector of $\mathscr{C}'$. We have

$$\begin{aligned}\operatorname{cov}((u + \alpha_0 v)^{\mathrm{T}} Y, w^{\mathrm{T}} Y) &= \operatorname{cov}(u^{\mathrm{T}} Y, w^{\mathrm{T}} Y) + \alpha_0 \operatorname{cov}(v^{\mathrm{T}} Y, w^{\mathrm{T}} Y) \\ &= \sigma^2 \left(u^{\mathrm{T}} w - \frac{u^{\mathrm{T}} v}{v^{\mathrm{T}} v} v^{\mathrm{T}} w \right).\end{aligned}$$

Note that $\mathscr{C} \cap \mathscr{C}' \subset \mathscr{C} \cap \mathscr{P} \subset \mathscr{P}$ and clearly, $\mathscr{C} \cap \mathscr{C}' \subset \mathscr{C}'$. Thus $\mathscr{C} \cap \mathscr{C}' \subset \mathscr{C}' \cap \mathscr{P}$. Hence $|\mathscr{C}' \cap \mathscr{P}| \geqslant |\mathscr{C} \cap \mathscr{C}'|$. Therefore $|u^{\mathrm{T}} w| \geqslant |v^{\mathrm{T}} w|$. Also, $|\mathscr{C}| > |\mathscr{P} \cap \mathscr{C}|$ and hence $v^{\mathrm{T}} v > |u^{\mathrm{T}} v|$. These two facts imply that

$$u^{\mathrm{T}} w - \frac{u^{\mathrm{T}} v}{v^{\mathrm{T}} v} v^{\mathrm{T}} w \neq 0.$$

Thus we may find a linear combination of $(u + \alpha_0 v)^{\mathrm{T}} Y$ and $w^{\mathrm{T}} Y$ which is unbiased for $\beta_i - \beta_j$ and has smaller variance than that of $(u + \alpha_0 v)^{\mathrm{T}} Y$. (To see this, just consider a linear combination $(u + \alpha_0 v)^{\mathrm{T}} Y + \gamma w^{\mathrm{T}} Y$ and minimize its variance with respect to γ. The fact that $w^{\mathrm{T}} \alpha_0 \neq 0$ ensures that the minimum is attained at $\gamma \neq 0$.) Then $(u + \alpha_0 v)^{\mathrm{T}} Y$ is not the BLUE of $\beta_i - \beta_j$ and (3) must be strict. □

The following result has been obtained by Klein and Randić [10] using concepts from electrical network theory.

Theorem 3. *Let G be a connected graph with $V = \{1, \ldots, n\}$, $E = \{e_1, \ldots, e_m\}$. Let L be the Laplacian and let $M = L^+$. Also, let $D = [d_{ij}]$ be the distance matrix of G. Then*

$$m_{ii} + m_{jj} - 2m_{ij} \leqslant d_{ij}, \quad i, j = 1, \ldots, n. \tag{6}$$

Furthermore, equality holds in (6) *if and only if there is a unique (i, j)-path in G.*

Proof. Consider the graph obtained by taking the disjoint union of G and a cycle $\mathscr{C}$. Let $\mathscr{P}$ be an (i, j)-path of minimum length. Then, using the notation of Theorem 2, $\lambda_{ij} = d_{ij}$ and $t_{ij} = 0$. Now the result follows immediately from Theorem 2. □

We remark that if H is *any* generalized inverse of L, then

$$m_{ii} + m_{jj} - 2m_{ij} = h_{ii} + h_{jj} - h_{ij} - h_{ji}.$$

To see this, set z^{ij} to be the vector

$$(0, \ldots, 0, 1, 0, \ldots, 0, -1, 0, \ldots, 0)^{\mathrm{T}},$$

where the 1 and the -1 occur at the ith place and the jth place, respectively. Then z^{ij} is in the column space of L and hence $z^{ij}L^{-}z^{ij}$ is invariant under the choice of generalized inverse. However, we use the Moore–Penrose inverse for convenience.

The following special case of Theorem 3 was observed in [2].

Corollary 4. *Let T be a tree with Laplacian matrix L and distance matrix D. Let $M = L^{+}$. Then for all i,j*

$$m_{ii} + m_{jj} - 2m_{ij} = d_{ij}.$$

The Wiener index, $W(G)$, of a graph G has been defined as

$$W(G) = \sum_{i<j} d_{ij},$$

and it has important applications in biochemistry, see [9,12]. Summing (6) with respect to i, j and keeping in mind that the row and column sums of M are zero, we get the following well-known fact – For any connected graph G with n vertices, $W(G) \geqslant n$ trace(L^{+}), and equality holds if and only if G is a tree. Recall that trace (L^{+}) is precisely the sum of the reciprocals of the nonzero eigenvalues of L.

Consider a block design in which v treatments are allocated in b blocks. We may associate a bipartite graph with $v + b$ vertices with the design, in which there are v vertices corresponding to the treatments and b vertices corresponding to the blocks. Two vertices are joined if one represents a treatment which appears in the block represented by the other vertex. Let n be the number of observations. Clearly, in order that the design be connected, the corresponding graph must be connected and this is true if and only if $n \geqslant v + b - 1$. Krafft and Schaefer [11] consider the situation, where $n = v + b$ (so the graph is unicyclic) and identify the designs which are A-optimal. In the course of their proof they obtain a special case of Theorem 2, see [11, Theorem 1, p. 377]. We have generalized their result to arbitrary (not necessarily bipartite) graphs. It appears from this connection that the linear model based on a graph as considered in the present section (see the proof of Theorem 2) can be a very useful tool in the area of optimality of block designs.

References

[1] R.B. Bapat, Linear Algebra and Linear Models, 2nd ed., Hindustan Book Agency, New Delhi, 1999.

[2] R.B. Bapat, Moore–Penrose inverse of the incidence matrix of a tree, Linear and Multilinear Algebra 42 (1997) 159–167.

[3] A. Ben-Israel, T.N.E. Greville, Generalized Inverses: Theory and Applications, Wiley-Interscience, New York, 1974.

[4] J.A. Bondy, U.S.R. Murty, Graph Theory with Applications, Macmillan, London, 1976.

[5] S.L. Campbell, C.D. Meyer Jr., Generalized Inverses of Linear Transformations, Dover, New York, 1991.

[6] P. Chebotarev, E. Shamis, The matrix-forest theorem and measuring relations in small social groups, Avtomatika i Telemekhanika 9 (1997) 124–136.

[7] M. Fiedler, Moore–Penrose involutions in the classes of Laplacians and simplices, Linear and Multilinear Algebra 39 (1995) 171–178.

[8] S. Kirkland, M. Neumann, B. Shader, Distances in weighted trees and group inverse of Laplacian matrices, SIAM J. Matrix Anal. Appl. 18 (1997) 827–841.

[9] D.J. Klein, Graph geometry, graph metrics and Weiner, Communications in Mathematical Match and in Computer Chemistry 35 (1997) 7–27.

[10] D.J. Klein, M. Randić, Resistance distance, J. Math. Chem. 12 (1993) 81–95.

[11] Krafft, Olaf, Schaefer, Martin, *A*-optimal connected block designs with nearly minimal number of observations, J. Statist. Plann. Inference 65 (1997) 375–386.

[12] R. Merris, An edge version of the matrix-tree theorem and the Wiener index, Linear and Multilinear Algebra 25 (1989) 291–296.

[13] R. Merris, Laplacian matrices of graphs: a survey, Linear Algebra Appl. 197 (1994) 143–176.

[14] J.W. Moon, On the adjoint of a matrix associated with trees, Linear and Multilinear Algebra 39 (1995) 191–194.

[15] C.R. Rao, Linear Statist. Inference Appl., Wiley, New York, 1973.