*J. Austral. Math. Soc.* **20** (Series B) (1977), 202–210

# TWO RELATED ESTIMATION PROBLEMS

W. BRISLEY

(Received 11 July 1977)

(Revised 19 October 1977)

**Abstract**

Two problems involving the "best" solution $X$ for a matrix equation $VXV^* = M$ are discussed, together with methods for their solution, and a generalization of one of the methods beyond matrix equations.

## 1. The two problems

Suppose $V$ and $M$ to be given $k \times n$ and $k \times k$ matrices respectively, and consider the equation

$$VXV^* = M \tag{1}$$

(where * indicates complex conjugate transpose—in general, the matrices will be taken over $C$).

In typical cases, this equation arises with $k < n$, and there is no diagonal solution $X$, the matrix $M$ being a "distortion" of a result derived from a supposed diagonal matrix, the "distortion" being due to noise or some other factor. For a particular case in signal processing, see, for example, [1].

Two problems arise:

**Problem (a).** *Select, from the set of all solutions $X$ of* (1), *that particular matrix $X$ for which the quantity*

$$\sum_{\substack{i \neq j \\ i=1 \ldots n \\ j=1 \ldots n}} |x_{ij}|^2$$

*is least.*

(If such a solution exists, we will term it the "tightest" solution of (1): one may reasonably describe it as being the solution which is "closest to being diagonal".

In typical cases, the matrix $V$ is usually arranged from experimental data, and is of rank $k$; in this circumstance (1) has solutions and so the problem makes sense. For a substantial discussion of further circumstances under which solutions may exist, see Rao [2].)

**Problem (b).** *Select, from the set of all diagonal $n \times n$ matrices, $D$, that particular one $D$ such that the quantity $|VDV^* - M|$ is least, where $|A|$ denotes*

$$\sum_{\substack{i=1\ldots k \\ i=j\ldots k}} |a_{ij}|^2.$$

(One may reasonably describe it as "the diagonal which is closest to giving $M$". Even if (1) has no solutions, this problem still makes sense, and then may be the most appropriate one for some signal processing estimation tasks (d'Assumpcao [1]).)

In each case, it is clear that a necessary condition for the problem to be well-posed is that the equation

$$VKV^* = 0 \tag{2}$$

admits no *diagonal* non-zero solutions—for, if "$K = D'$" was such a solution, then whenever "$X = S$" solved (a), then so would "$X = S + D'$", and, whenever "$D = B$" solved (b), then so would "$D = B + D'$". Consequently, in assuming that the problems are well-posed, we assume that

$$\text{If } VKV^* = 0_{k\times k} \text{ and } K \text{ is diagonal, then } K = 0_{n\times n}. \tag{3}$$

(We shall return to some analysis of the conditions under which (3) holds, in Section 4.)

## 2. Solution of Problem (a)

We assume that (1) does not lack solutions, and adopt the point of view that (1), (2) and (3) are referring to the linear map $\phi$ from the space of $n \times n$ matrices to the space of $k \times k$ matrices, where

$$\phi(A) = VAV^*.$$

Consequently, the solution set of (2) is the kernel of $\phi$, statement (3) is a property of that kernel, and we are considering a search amongst all matrices of the form

$$X = P + K,$$

where $P$ is a *particular* matrix such that $\phi(P) = M$, and $K$ is such that $\phi(K) = 0$.

Let $G$ be any "generalized inverse" for $V$; that is, let $G$ be one of the many $n \times k$ matrices such that

$$VGV = V \tag{4}$$

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

(and hence also

$$V^*G^*V^* = V^*). \tag{4'}$$

The matrix $G$ has only to satisfy (4) for the following, so any of the "generalized inverses" defined in the literature will suffice (see, for example, Rao [3], pp. 24–26 and also Noble and Daniel [4]). The equation (4) implies that

$$V\mathbf{k} = \mathbf{0} \quad \text{if and only if } \mathbf{k} = (I - GV)\mathbf{r} \tag{4''}$$

for some arbitrary vector $\mathbf{r}$ and

$$\text{if } V\mathbf{x} = \mathbf{b} \text{ has solutions, then}$$

$$\mathbf{x} = G\mathbf{b} \text{ is one of them.} \tag{4'''}$$

It follows from (4‴) that $GMG^*$ is a particular solution for (1). Letting $K$ denote any solution of (2) (that is, $K \in \ker \phi$), we have from (4″) that $KV^* = (I - GV)A$ for some arbitrary $n \times k$ matrix $A$, and hence (using (4″) again) that

$$K = (I - GV)AG^* + B(I - V^*G^*), \tag{5}$$

where $A$ and $B$ can be any $n \times k$ and $n \times n$ matrices. Adopting $P = GMG^*$, we need only select $K$, which we do using a Fourier technique, as follows.

Suppose $V$ has rank $r$. Then the subspace of $C^n$ $\{\mathbf{k} \mid V\mathbf{k} = \mathbf{0}\}$ has a basis $\{\mathbf{k}_1, \mathbf{k}_2, \ldots, \mathbf{k}_\nu\}$ with $\nu = n - r$ (note that a suitable basis will have already appeared in the calculation of $G$). Let $C_{ij}$ denote the $n \times n$ matrix which has $\mathbf{k}_j$ entered as column $i$, and zero for all other entries. Then clearly any matrix of the form

$$\sum_{\substack{i=1\ldots n \\ j=1\ldots \nu}} \alpha_{ij} C_{ij} + \sum_{\substack{i=1\ldots n \\ j=1\ldots \nu}} \beta_{ij} C_{ij}^* \tag{6}$$

will be a solution of (2). Since (5) displays each such $K$ as a sum

$$K = R + S \quad \text{with } VR = 0 \quad \text{and} \quad SV^* = 0 \tag{5'}$$

it follows that $\ker \phi$ is spanned by the set $\mathscr{K} \cup \mathscr{K}^*$, where

$$\mathscr{K} = \{C_{ij}\}_{\substack{i=1\ldots n \\ j=1\ldots \nu}}, \quad \mathscr{K}^* = \{C_{ij}^*\}_{\substack{i=1\ldots n \\ j=1\ldots \nu}}.$$

On the space of all $n \times n$ matrices, set up the inner product $\langle ; \rangle$ by *defining* that

$$\langle A; B \rangle = \sum_{i \neq j} a_{ij} \overline{b_{ij}}. \tag{7}$$

It is clear that this *is* an inner product, that it is not positive definite, but also that:

*when restricted to* $\ker \phi$, *under assumption* (3), $\langle ; \rangle$ *is a positive definite inner product on* $\ker \phi$.

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

One can then define $\|X\|$ to mean $\sqrt{(\langle X; X\rangle)}$ and note that $\|X\|$ is 0 if and only if $X$ is diagonal.

Although $\mathscr{K} \cup \mathscr{K}^*$ is not necessarily a linearly independent set, it will yield (by, for example, a modified Gram–Schmidt process) an orthonormal basis for $\ker \phi$, say

$$\{K_1, K_2, \ldots, K_m\} \quad \text{with} \quad \langle K_i, K_j\rangle = \delta_{ij}, \quad m \leqslant 2nv.$$

Then any solution of (1) reads

$$\text{“}X = S \quad \text{where} \quad S = P + \sum_{i=1}^{m} \alpha_i K_i\text{”}$$

and Problem (a) is reduced to selecting the $\alpha_i$ to minimize $\|S\|$. Since $P$ has been adopted already, and hence the numbers $\langle P; P\rangle$ and $f_i = \langle P; K_i\rangle$ are fixed, we calculate that

$$\langle S; S\rangle = \langle P; P\rangle + \sum_{1}^{m} \alpha_i \bar{f}_i + \sum_{1}^{m} \bar{\alpha}_i f_i + \sum_{1}^{m} \alpha_i \bar{\alpha}_i,$$

$$= \langle P; P\rangle + \sum_{i=1}^{m} (\alpha_i + f_i)(\bar{\alpha}_i + \bar{f}_i) - \sum_{i=1}^{m} f_i \bar{f}_i.$$

Thus, a minimum for $\|S\|$ is achieved if and only if each $\alpha_i$ is $-f_i$, and the solution to problem (*a*) appears as

$$\text{“}X = S \quad \text{with} \quad S = P - \sum_{i=1}^{m} \langle P; K_i\rangle K_i\text{”}. \tag{8}$$

(That the result is independent of the particular choice of $P$ is clear, since if $S_j = P_j - \sum_{i=1}^{m} \langle P_j; K_i\rangle K_i$ $(j = 1, 2)$ then, since $P_1 - P_2 \in \ker \phi$, and hence

$$P_1 - P_2 = \sum_{i=1}^{m} \langle (P_1 - P_2); K_i\rangle K_i,$$

we have $S_1 - S_2 = 0$. It is equally “standard” that the result is independent of the particular orthonormal basis chosen for $\ker \phi$. Thus, it does make sense to speak of *the* tightest solution $X = S$ given by (8).)

Some further remarks are of interest:

(i) *If $M$ of* (1) *is Hermitian, then the “tightest” solution is also Hermitian.*

This can be seen as follows: from (6), we have

$$S^* = P^* - \sum_{i=1}^{m} \langle \overline{P; K_i}\rangle K_i^*.$$

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

But since $\overline{\langle A; B\rangle} = \langle A^*; B^*\rangle$, the set $\{K_1^*, \ldots, K_n^*\}$ is also an orthonormal basis for $\ker\phi$, and since $P^*$ is also a particular solution of (1) when $M$ is Hermitian, then $S^*$ is merely the "tightest solution" using $P^*$ and $\{K_1^*, \ldots, K_n^*\}$. Hence $S = S^*$.

(ii) *It can be arranged that each $K_i$ is Hermitian, and hence, if $M$ is Hermitian, that each $\langle P; K_i\rangle$ be real.*

We note that the Gram–Schmidt process preserves "Hermitian", since it only involves calculations such as $A/\langle A; A\rangle$ and $A - \langle A; B\rangle B$, and $\langle A; B\rangle$ is real if both $A$ and $B$ are Hermitian. If then, we use, instead of $\mathscr{K} \cup \mathscr{K}^*$, the spanning set

$$\{(C_{ij} + C_{ij}^*)\}_{\substack{i=1\ldots n\\ j=1\ldots v}} \cup \{i(C_{ls} - C_{ij}^*)\}_{\substack{l=1\ldots n\\ s=1\ldots v}}$$

(which consists entirely of Hermitian matrices) and apply the Gram–Schmidt process, the resulting $\{K_1, \ldots, K_n\}$ will consist entirely of Hermitians. If, further, $M = M^*$, and we use $P = GMG^*$, then $P^* = P$, and hence

$$\overline{\langle P; K_i\rangle} = \langle P^*; K_i^*\rangle = \langle P; K_i\rangle.$$

(iii) *In the special case that $V$, $M$ are real, and $M$ is symmetric, we need only consider real symmetric matrices, in which case the dimension of* $\ker\phi$ *is no greater than $nv$.*

For, by remark (ii), we have collapse from "Hermitian" to "symmetric" wherever appropriate. If $K$ is in $\ker\phi$, it can be written as *some* linear combination

$$K = \sum_{\substack{i=1\ldots n\\ j=1\ldots v}} \alpha_{ij} C_{ij} + \sum_{\substack{i=1\ldots n\\ j=1\ldots v}} \beta_{ij} C_{ij}^{\mathrm{T}}.$$

But also $K = K^{\mathrm{T}} = \sum \beta_{ij} C_{ij} + \sum \alpha_{ij} C_{ij}^{\mathrm{T}}$ whence $K = \sum \gamma_{ij}(C_{ij} + C_{ij}^{\mathrm{T}})$ where $\gamma_{ij} = \frac{1}{2}(\alpha_{ij} + \beta_{ij})$. Hence, we may take $\{(C_{ij} + C_{ij}^{\mathrm{T}})\}_{\substack{i=1\ldots n\\ j=1\ldots v}}$ as a spanning set. That dimension $nv$ *can* be attained is demonstrated by the following example.

$$V = \begin{bmatrix} 1 & 0 & 0 & -1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & -1 \end{bmatrix}; \quad v = 1, \quad \mathbf{k}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

giving

$$K_1 = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 & 1 & 1 & 1 \\ 1 & & & \\ 1 & & 0 & \\ 1 & & & \end{bmatrix}, \quad K_2 = \frac{1}{4\sqrt{3}} \begin{bmatrix} -2 & 2 & -1 & -1 \\ 2 & 6 & 3 & 3 \\ -1 & 3 & 0 & 0 \\ -1 & 3 & 0 & 0 \end{bmatrix},$$

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

$$K_3 = \frac{1}{4\sqrt{5}}\begin{bmatrix} -2 & -2 & 3 & -1 \\ -2 & -2 & 3 & -1 \\ 3 & 3 & 8 & 4 \\ -1 & -1 & 4 & 0 \end{bmatrix}, \quad K_4 = \frac{1}{\sqrt{30}}\begin{bmatrix} -1 & -1 & -1 & 2 \\ -1 & -1 & -1 & 2 \\ -1 & -1 & -1 & 2 \\ 2 & 2 & 2 & 5 \end{bmatrix}$$

giving, when

$$M = \begin{bmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix},$$

the "tightest solution" for Problem (a) as

$$S = \tfrac{1}{6}\begin{bmatrix} 10 & 1 & -2 & 1 \\ 1 & 4 & 1 & -2 \\ -2 & 1 & 10 & 1 \\ 1 & -2 & 1 & 4 \end{bmatrix}.$$

### 3. Solution of Problem (b)

Here, we take the view that $V$ affords a linear map $\psi$ from the space of all *diagonal* $n \times n$ matrices into the space of all $k \times k$ matrices, by

$$\psi(D) = VDV^*. \tag{9}$$

Condition (3) assures us that $\psi$ is of rank $n$, and Problem (b) is solved by finding that matrix $Q$ in Image $\psi$ which is "closest" to $M$, and then setting $D = \psi^{-1}(Q)$. In this case, the relevant inner product on the space of $k \times k$ matrices is

$$\langle\langle A; B\rangle\rangle \text{ defined as } \sum_{\substack{i=1\ldots k \\ j=1\ldots k}} a_{ij}\bar{b}_{ij}$$

(so that the corresponding norm $|\cdot|$ is as in Problem (b)). Calculating

$$\{\psi(D_1), \psi(D_2), \ldots, \psi(D_n)\}$$

for an appropriate linearly independent set of diagonals gives a basis for Image $\psi$, hence an orthonormal basis for Image $\psi$, say $\{B_i\}_{i=1,\ldots,n}$, and then $Q$ is determined uniquely as

$$Q = \sum_{i=1}^{n} \langle\langle M; B_i\rangle\rangle B_i.$$

(By arguments similar to previous ones, we may take the $B_i$ to be Hermitian, and so if $M$ is Hermitian, all the $\langle\langle M; B_i\rangle\rangle$ are real, and $Q$ is Hermitian.)

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

The equation

$$\psi(D) = Q \tag{10}$$

can be rewritten as

$$N\mathbf{d} = \mathbf{q}, \tag{10'}$$

where $\mathbf{d}$ is the diagonal of $D$, $\mathbf{q}$ is a suitable "unwrapping" of $Q$ and $N$ is $k^2 \times n$. Since $N$ is of rank $n$ by (3), we have $\mathbf{d} = Z\mathbf{q}$ whenever $Z$ is a generalized inverse for $N$. Equation (10) and condition (3) are discussed below.

As a matter of comparison with Problem (a), the process applied to the same $V, M$ in the illustration given yields

$$\left\{ B_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad B_3 = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad B_4 = \frac{1}{\sqrt{6}} \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \right\}$$

and hence

$$D = \tfrac{1}{3} \begin{bmatrix} 4 & & & 0 \\ & 4 & & \\ & & 4 & \\ 0 & & & 2 \end{bmatrix}$$

as a solution for Problem (b).

## 4. The "well-posing" of the problems

Condition (3) is clearly necessary for the problems to be well-posed. Since, with it, they each have unique solutions, it is also *sufficient* for them to be well-posed, as long as, in Problem (a), we assume the existence of at least one solution $X$ to equation (1). The equations (10) and (10') (and hence condition (3)) can be explicitly written in numerous ways: one such is to "unwrap" $Q$ as the column whose entries read in order

$$q_{11}, q_{12}, \ldots, q_{1k}, q_{21}, \ldots, q_{2k}, q_{31}, \ldots, q_{k1}, \ldots, q_{kk}.$$

In this case (the diagonal $d$ reading $d_{11}, d_{22}, \ldots, d_{nn}$) one has $N$ consisting of $k^2$ rows, the $i$th block of $k$ rows reading:

$$\begin{aligned} &v_{i1}\bar{v}_{11}, v_{i2}v_{12}, \ldots, v_{in}\bar{v}_{1n}, \\ &v_{i1}\bar{v}_{21}, v_{i2}\bar{v}_{22}, \ldots, v_{in}\bar{v}_{2n}, \\ &\quad\vdots \\ &v_{i1}\bar{v}_{k1}, v_{i2}\bar{v}_{k2}, \ldots, v_{in}\bar{v}_{kn}. \end{aligned}$$

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

For condition (3), that is, for $N$ to be rank $n$, we clearly need at least, that $n \leqslant k^2$. (If $N$ is of rank $n$, with $n = k^2$, then equation (1) has a unique *diagonal* solution, and both problems become, trivially, "solve $N\mathbf{d} = \mathbf{m}$ where $\mathbf{m}$ is $M$ unwrapped".) If $M$ is Hermitian, then the $Q$ of (10) is Hermitian, and so $\psi(D) = Q$ implies that $D$ must be real; since the rows of $N$ occur in complex conjugate pairs, this leaves just $\frac{1}{2}k(k+1)$ equations to solve, $k$ of them being real: although this does not alter the bound on the necessary rank of $N$, it does provide a smaller matrix than $N$ for the computation of the appropriate generalized inverse.

Two special cases are worthy of mention. In at least one type of application, each entry in $V$ is of modulus 1, so that $N$ has at least $k$ rows with each entry 1. In this case, then, one has "$k^2-k+1 \geqslant n$" as a necessary condition for each of the problems to be well-posed, and $M$ *not* having equal diagonal entries is then a sufficient condition for there to be no diagonal solution to equation (1).

In applications where all matrices are real, there will be at most $\frac{1}{2}k(k+1)$ distinct rows in $N$, so a necessary condition for the problems to be well-posed is that $n \leqslant \frac{1}{2}k(k+1)$. If, further, $M$ is symmetric, $n$ being $\frac{1}{2}k(k+1)$ will render both problems trivial, in that (1) will then have a unique diagonal solution.

Further discussion of the relations between the rank of $N$ and properties of $V$ can be found in [2].

## 5. Generalization of Problem (a) and its solution

The Problem (a) is a special case of the following: *given the linear map* $\phi$: $U \to V$, *with* $S$ *a subspace of the vector space* $U$, *and* $\mathbf{g}$ *in* $V$, *find that solution* $\mathbf{x}$ *of* $\phi(\mathbf{x}) = \mathbf{g}$ *such that* $\mathbf{x}$ *is "closest possible" to being a member of* $S$.

If we take it that $\langle\,;\,\rangle$ is a positive definite inner product on $U$, so that $U$ splits as $S \oplus S^{\perp}$ (each $\mathbf{u}$ in $U$ being written $\mathbf{u} = s(\mathbf{u}) + r(\mathbf{u})$ uniquely, with $s(\mathbf{u}) \in S$, $r(\mathbf{u}) \in S^{\perp}$), the obvious measure is to define

$$\langle\langle \mathbf{u}; \mathbf{v} \rangle\rangle^{\cdot} = \langle r(\mathbf{u}); r(\mathbf{v}) \rangle,$$

so that $\langle\langle\,;\,\rangle\rangle$ is a positive definite inner product on the factor space $U/S$ and, also, on $K/K \cap S$ where $K$ is the kernel of $\phi$. Taking $\{k_1, k_2, \ldots\}$ (modulo $K \cap S$) as an orthonormal basis for $K/K \cap S$, with respect to $\langle\langle\,;\,\rangle\rangle$, it follows (as in the particular case of Section 2) that of all possible solutions of $\phi(\mathbf{x}) = \mathbf{g}$, that solution with $\langle\langle \mathbf{x}; \mathbf{x} \rangle\rangle$ least possible is

$$\mathbf{x} = \mathbf{p} - \sum \langle\langle \mathbf{p}; \mathbf{k}_i \rangle\rangle \mathbf{k}_i \quad \text{(modulo } K \cap S\text{)},$$

where $\mathbf{p}$ is any particular solution to $\phi(\mathbf{p}) = \mathbf{g}$.

https://doi.org/10.1017/S0334270000001570 Published online by Cambridge University Press

Perhaps the simplest examples (other than the more obvious signal processing-parameter estimation problems) are in control problems; as two samples which are simple enough for scratch pad checking, we have:

(i) *Find the input* $y(t)$ *with least R.M.S. over* $0 \leqslant t \leqslant 1$, *such that* $y - y' = t^2 - 2t$. (Here $S$ is zero, $K$ is spanned by the single function $e^t$, and the required function is $y(t) = t^2 - (2(e-2)/(e^2-1))\,e^t$.)

(ii) *Find the input* $y(t)$, *which is closest to being linear such that* $y - y' = t^2 - 2t$ in the interval $0 \leqslant t \leqslant 1$. (Here one takes $S$ as the subspace of all linear functions, with $\langle f; g \rangle$ being $\int_0^1 fg\,dt$, so that the required function is $y(t) = t^2 - (1/3(3-e))\,e^t$.)

The usefulness of the method, in this context, increases with the degree of the relevant differential equation, since it merely involves the projection of a new particular solution on to the fixed space $K/K \cap S$ each time a "new" target function is presented, rather than re-solving a "new" minimization problem. In practice, the parameters needed are immediately available by quadrature, whereas the corresponding *ab initio* minimization problem involves not only quadratures but also the solution of a set of simultaneous equations which are often, because of their source, fairly badly conditioned.

**Acknowledgement**

The author wishes to thank the referee for his comments and suggestions.

**REFERENCES**

[1] H. A. d'Assumpcao, "Estimation of sound directionality", *Australian Weapons Research Establishment* WRE–TN–1369 (WR & D).

[2] C. R. Rao, "Estimation of heteroscedastic variance", *J. Amer. Statist. Ass.* 65 (1970), 161–172.

[3] C. R. Rao, *Linear statistical inference* (John Wiley & Sons Inc., New York 1965).

[4] B. Noble and J. W. Daniel, *Applied linear algebra* (Prentice Hall, New York 2nd ed., 1977).

Department of Mathematics
University of Newcastle
New South Wales 2308
Australia