A note on subset selection for matrices

A Note on Subset Selection for Matrices

F.R. de Hoog, CSIRO Mathematical and Information Sciences
PO Box 664, Canberra, ACT 2601, Australia.
email: Frank.deHoog@csiro.au

R.M.M. Mattheij, Department of Mathematics and Computer Science,
Technische Universiteit Eindhoven, PO Box 513, 5600 MB Eindhoven, The Netherlands.
email: r.m.m.mattheij@tue.nl

Abstract

In an earlier papers the authors established a result to select subsets of a matrix that are as "non-singular" as possible in a numerical sense. The major result was not constructive. In this note we give a constructive proof and moreover a sharper bound.

1. Introduction

In [2] the problem of selecting k rows from an $m \times n$ matrix such that the resulting matrix was as non-singular as possible was examined. That is, for $\mathbf{X} \in \boldsymbol{R}^{m \times n}$ find a permutation matrix $\mathbf{P} \in \boldsymbol{R}^{m \times m}$ so that

$$\mathbf{PX} = \begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}, \quad \mathbf{A} \in \boldsymbol{R}^{k \times n}, \tag{1}$$

where A is the matrix in question, $m, k > n$ and $\operatorname{rank}(\mathbf{X}) = n$.

To motivate this problem, consider the problem of regression where we have a vector of n observations

$$\mathbf{y} = \mathbf{A}\boldsymbol{\theta} + \boldsymbol{\delta},$$

where $\mathbf{A} \in \boldsymbol{R}^{k \times n}$ is a design matrix whose rows are a subset of the rows of $\mathbf{X} \in \boldsymbol{R}^{m \times n}$, $\boldsymbol{\theta} \in \boldsymbol{R}^{n}$ is a vector of unknown parameters that is to be determined and $\boldsymbol{\delta} \in \boldsymbol{R}^{k}$ is a vector whose components are independent and identically normally distributed. Such problems occur when observations are expensive and only a subset of all possible measurements is feasible. The least squares estimate of the unknown parameters is $\hat{\boldsymbol{\theta}} = \mathbf{A}^{+}\mathbf{y}$ where $\mathbf{A}^{+}$ is the Moore-Penrose inverse. For a given design matrix $\mathbf{A}$ and confidence coefficient, the confidence ellipsoid for $\boldsymbol{\theta}$ is given by $\left\{ \boldsymbol{\theta} \,\middle|\, (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^{T} \mathbf{A}^{T}\mathbf{A}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \le \text{constant} \right\}$. The content of this ellipsoid is proportional to $(\det \mathbf{A}^{T}\mathbf{A})^{-\frac{1}{2}}$ and it is natural to make this as small as possible. That is, we choose the design matrix $\mathbf{A}$ to maximise $\det \mathbf{A}^{T}\mathbf{A}$. Such designs are called D-optimal designs (see Silvey [5] for a more detailed discussion). However, optimality will depend on the application. For example minimising

$\left\|\mathbf{A}^{+}\right\|_F = \sqrt{\text{Trace}\left(\mathbf{A}^T\mathbf{A}\right)^{-1}}$ ensures that the expected mean squared error of $\boldsymbol{\theta}$ is minimised. E-optimal designs (see Silvey [5]) maximise the smallest singular value of $\mathbf{A}$ (or equivalently, maximise $\left\|\mathbf{A}^{+}\right\|_2 = \left\|\left(\mathbf{A}^T\mathbf{A}\right)^{-1}\right\|_2^{\frac{1}{2}}$). Further applications are described in [2]

Row selection is often implemented using a QR decomposition of X^T with column interchange to maximize the size of the pivots (see [1] and also [3], section 12.2). This algorithm usually works well but there are examples [4, p31] where the pivot size does not adequately reflect the size of the singular values. As a consequence bounds from the analysis of such algorithms would lead to poor bounds for the singular values and related quantities such as $\det \mathbf{A}^T\mathbf{A}$.

In [2] the present authors derived upper bound for $\left\|\mathbf{A}^{+}\right\|_F$ and the singular values of $\mathbf{A}$. In this note, we extend these results by deriving a constructive derivation for the bounds on the singular values and new lower bounds for $\det \mathbf{A}^T\mathbf{A}$.
In section 2 we give the main results and in particular a sharper bound for $\left\|\mathbf{A}^{+}\right\|_F$. In section 3 we show that this bound is sharper than the one obtained earlier, at least asymptotically..

2. Results

We can rewrite (1) as

$$\mathbf{PX} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = \begin{bmatrix}\mathbf{Q}\\ \mathbf{Y}\end{bmatrix}\left(\mathbf{A}^T\mathbf{A}\right)^{\frac{1}{2}} \qquad (2)$$

where

$$\mathbf{Q} := \mathbf{A}\left(\mathbf{A}^T\mathbf{A}\right)^{-\frac{1}{2}}$$

$$\mathbf{Y} := \mathbf{B}\left(\mathbf{A}^T\mathbf{A}\right)^{-\frac{1}{2}}$$

It follows that,

$$\mathbf{X}^T\mathbf{X} = \left(\mathbf{A}^T\mathbf{A}\right)^{\frac{1}{2}}\left(\mathbf{I} + \mathbf{Y}^T\mathbf{Y}\right)\left(\mathbf{A}^T\mathbf{A}\right)^{\frac{1}{2}} \qquad (3)$$

Thus

$$\det\left(\mathbf{X}^T\mathbf{X}\right) = \det\left(\mathbf{I} + \mathbf{Y}^T\mathbf{Y}\right)\det\left(\mathbf{A}^T\mathbf{A}\right) \qquad (4)$$

and so maximising $\det \mathbf{A}^T\mathbf{A}$ is equivalent to minimizing $\det\left(\mathbf{I}+\mathbf{Y}^T\mathbf{Y}\right)$. From the arithmetic-geometric inequality we have

$$\det\left(\mathbf{I}+\mathbf{Y}^T\mathbf{Y}\right) \le \left(1+\tfrac{1}{n}\|\mathbf{Y}\|_F^2\right)^n \qquad (5)$$

This suggests that when $\mathbf{P}$ is chosen so that $\mathbf{Y}$ is not too large, then $\det \mathbf{A}^T\mathbf{A}$ will not be small. By applying the usual variational formulation for singular values to (3), we obtain

$$\sigma_l^2(\mathbf{A}) \le \sigma_l^2(\mathbf{X}) \le \left(1+\|\mathbf{Y}\|_2^2\right)\sigma_l^2(\mathbf{A}),\ l=1,\cdots,\mathrm{n} \qquad (6)$$

where $\sigma_l(\mathbf{A})$ and $\sigma_l(\mathbf{X})$ are the singular values of $\mathbf{A}$ and $\mathbf{X}$ respectively. Thus, the singular values of $\mathbf{X}$ will not be small if $\|\mathbf{Y}\|_2$ is not large.

We now show that a permutation exists so that the matrix $\mathbf{Y}$ that is not large. This result was established in [2] by assuming that $\mathbf{P}$ was chosen to maximise $\det \mathbf{A}^T\mathbf{A}$; the poof, however, as not constructive. In the present note, we give a construction based on a greedy algorithm where rows of $\mathbf{X}$ are deleted, one at a time, so as to minimise the Frobenius norm of $\mathbf{Y}$ at each step.

Theorem 1. There is a permutation matrix $\mathbf{P}$ so that (2) holds with

$$\|\mathbf{Y}\|_F^2 \le \frac{(m-k)n}{k-n+1}.$$

Proof: Let

$$\mathbf{A}=\begin{bmatrix}\mathbf{a}_1^T\\ \vdots\\ \mathbf{a}_k^T\end{bmatrix},\quad \mathbf{Q}=\begin{bmatrix}\mathbf{q}_1^T\\ \vdots\\ \mathbf{q}_k^T\end{bmatrix},\quad \mathbf{Y}=\begin{bmatrix}\mathbf{y}_1^T\\ \vdots\\ \mathbf{y}_{m-k}^T\end{bmatrix},$$

and note that the columns of $\mathbf{Q}$ are orthogonal. Indeed,

$$\sum_{r=1}^{k}\mathbf{q}_r\mathbf{q}_r^T = \mathbf{Q}^T\mathbf{Q}=\mathbf{I},$$

and .

$$\|\mathbf{q}_r\|_2 \le 1.$$

Suppose $k>n$ and that we wish to delete a row of $\mathbf{A}$. We define

$$\mathbf{A}_j := \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_{j-1}^T \\ \mathbf{a}_{j+1}^T \\ \vdots \\ \mathbf{a}_k^T \end{bmatrix}, \quad \mathbf{B}_j := \begin{bmatrix} \mathbf{a}_j^T \\ \hline \mathbf{B} \end{bmatrix},$$

and can then write for some permutation matrix $\tilde{\mathbf{P}}_j$

$$\tilde{\mathbf{P}}_j \mathbf{X} = \begin{bmatrix} \mathbf{A}_j \\ \mathbf{B}_j \end{bmatrix}, \quad \mathbf{A}_j \in \boldsymbol{R}^{(\text{k-1}) \times \text{n}}, \ \mathbf{B}_j \in \boldsymbol{R}^{(\text{m-k+1}) \times \text{n}}.$$

From this it follows that

$$\tilde{\mathbf{P}}_j \mathbf{X} = \begin{bmatrix} \mathbf{A}_j \\ \mathbf{B}_j \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_j \\ \mathbf{Y}_j \end{bmatrix} \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{\frac{1}{2}},$$

where

$$\mathbf{Q}_j = \mathbf{A}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-\frac{1}{2}}, \quad \mathbf{Y}_j = \mathbf{B}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-\frac{1}{2}}.$$

We have

$$\begin{aligned} \left\| \mathbf{Y}_j \right\|_F^2 &= \text{Trace}\left(\left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-1/2} \mathbf{B}_j^T \mathbf{B}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-1/2} \right) \\ &= \text{Trace}\left(\mathbf{B}_j^T \mathbf{B}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-1} \right) \\ &= \text{Trace}\left(\left(\mathbf{B}^T \mathbf{B} + \mathbf{a}_j \mathbf{a}_j^T \right) \left(\mathbf{A}^T \mathbf{A} - \mathbf{a}_j \mathbf{a}_j^T \right)^{-1} \right) \\ &= \text{Trace}\left(\left(\mathbf{Y}^T \mathbf{Y} + \mathbf{q}_j \mathbf{q}_j^T \right) \left(\mathbf{I} - \mathbf{q}_j \mathbf{q}_j^T \right)^{-1} \right) \\ &= \left\| \mathbf{Y} \right\|_F^2 + \frac{1}{1 - \left\| \mathbf{q}_j \right\|_2^2} \left(\left\| \mathbf{Y} \mathbf{q}_j \right\|_2^2 + \left\| \mathbf{q}_j \right\|_2^2 \right) \end{aligned}$$

Now let $\left\| \boldsymbol{Y}_j \right\|_F$ be minimized when $p = j$. Then,

$$\left(1 - \left\| \mathbf{q}_j \right\|_2^2 \right) \left\| \mathbf{Y}_p \right\|_F^2 \le \left(1 - \left\| \mathbf{q}_j \right\|_2^2 \right) \left\| \mathbf{Y} \right\|_F^2 + \left(\left\| \mathbf{Y} \mathbf{q}_j \right\|_2^2 + \left\| \mathbf{q}_j \right\|_2^2 \right).$$

On summing over j and noting that

$$\sum_{j=1}^{k} \left\| \mathbf{q}_j \right\|_2^2 = n,$$
$$\sum_{j=1}^{k} \left\| \mathbf{Y}\mathbf{q}_j \right\|_2^2 = \left\| \mathbf{Y} \right\|_F^2,$$

we obtain

$$(k-n)\left\| \boldsymbol{Y}_p \right\|_F^2 \le (k-n+1)\left\| \boldsymbol{Y} \right\|_F^2 + n. \qquad (7)$$

We can use this construction, starting with $\boldsymbol{X}$ and then deleting a row at the time whilst insuring that $\left\| \boldsymbol{H} \right\|_F = \left\| \boldsymbol{Y} \right\|_F$ is minimised at each step to construct

$$\boldsymbol{PX} = \begin{bmatrix} \boldsymbol{A} \\ \boldsymbol{B} \end{bmatrix}, \;\; \boldsymbol{A} \in \boldsymbol{R}^{\mathrm{k \times n}}, \;\; \boldsymbol{B} \in \boldsymbol{R}^{\mathrm{(m\text{-}k) \times n}}, \;\; k > n.$$

From (7) it follows by induction that such a construction satisfies

$$\left\| \mathbf{Y} \right\|_F^2 \le \frac{(m-k)n}{k-n+1}. \qquad \#$$

Theorem 1 and (4), (5) imply:

Corollary 1 There is a permutation matrix $\mathbf{P}$ so that

$$\det\left(\mathbf{A}^T\mathbf{A}\right) \ge \det\left(\mathbf{X}^T\mathbf{X}\right)\left(\frac{k-n+1}{m-n+1}\right)^n. \qquad (8)$$

In [2, cf Theorem 2], a greedy algorithm was presented where rows of $\mathbf{X}$ are deleted, one at a time, so as to minimise the Frobenius norm of $\mathbf{A}^+$ at each step was, which read

Theorem 2 There is a permutation matrix $\mathbf{P} \in \boldsymbol{R}^{m \times m}$ such that (1) holds with

$$\left\| \mathbf{A}^+ \right\|_F^2 \le \frac{m-n+1}{k-n+1}\left\| \mathbf{X}^+ \right\|_F^2.$$

This theorem can also be used to give an alternative proof of Corollary 1 (7)

Proof (of Corollary 1, alternative)**:** If we apply Theorem 2 to $\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-\frac{1}{2}}$, there is a permutation matrix $\mathbf{P}$ such that

$$\mathbf{PX}\left(\mathbf{X}^T\mathbf{X}\right)^{-\frac{1}{2}} = \begin{bmatrix}\mathbf{W}\\ \mathbf{Z}\end{bmatrix}, \quad \mathbf{W} \in \boldsymbol{R}^{\mathrm{k}\times\mathrm{n}},$$

with

$$\text{Trace}\left(\mathbf{W}^T\mathbf{W}\right)^{-1} = \left\|\mathbf{W}^+\right\|_F^2 \le \frac{m-n+1}{k-n+1}\left\|\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-\frac{1}{2}}\right\|_F^2 = n\left(\frac{m-n+1}{k-n+1}\right).$$

Moreover,

$$\mathbf{PX} = \begin{bmatrix}\mathbf{W}\\ \mathbf{Z}\end{bmatrix}\left(\mathbf{X}^T\mathbf{X}\right)^{\frac{1}{2}} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix},$$

and hence

$$\det\left(\mathbf{A}^T\mathbf{A}\right) = \det\left(\mathbf{W}^T\mathbf{W}\right)\det\left(\mathbf{X}^T\mathbf{X}\right) . \qquad (9).$$

From the geometric-arithmetic mean inequality, we have

$$\det\left(\mathbf{W}^T\mathbf{W}\right)^{-1} \le \left(\tfrac{1}{n}\text{Trace}\left(\mathbf{W}^T\mathbf{W}\right)^{-1}\right)^n \le \left(\frac{m-n+1}{k-n+1}\right)^n,$$

and the result follows on substitution of this inequality in (9). #

The bound for $\det\mathbf{A}^T\mathbf{A}$ in corollary 1 follows from bounds on $\left\|\mathbf{Y}\right\|_F$, and proof above on $\left\|\mathbf{A}^+\right\|_F$ respectively. A somewhat tighter bound can be obtained by analysing a greedy algorithm where $\det\mathbf{A}^T\mathbf{A}$ is maximized at each step.

Theorem 3 There is a permutation matrix $\mathbf{P} \in \boldsymbol{R}^{m\times m}$ such that (1) holds with

$$\det\left(\mathbf{A}^T\mathbf{A}\right) \ge \det\left(\mathbf{X}^T\mathbf{X}\right)\prod_{j=k+1}^{m}\frac{j-n}{j} = \frac{k!(m-n)!}{m!(k-n)!}\det\left(\mathbf{X}^T\mathbf{X}\right) . \qquad (10)$$

Proof: As in theorem 1, we have

$$\mathbf{PX} = \begin{bmatrix}\mathbf{A}\\ \mathbf{B}\end{bmatrix} = \begin{bmatrix}\mathbf{Q}\\ \mathbf{Y}\end{bmatrix}\left(\mathbf{A}^T\mathbf{A}\right)^{\frac{1}{2}},$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_k^T \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} \mathbf{q}_1^T \\ \vdots \\ \mathbf{q}_k^T \end{bmatrix}, \ \mathbf{Y} = \begin{bmatrix} \mathbf{y}_1^T \\ \vdots \\ \mathbf{y}_{m-k}^T \end{bmatrix},$$

and the columns of $\mathbf{Q}$ are orthogonal.

Suppose $k > n$ and that we wish to delete a row of $\mathbf{A}$. We define

$$\mathbf{A}_j = \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_{j-1}^T \\ \mathbf{a}_{j+1}^T \\ \vdots \\ \mathbf{a}_k^T \end{bmatrix}, \ \mathbf{B}_j = \begin{bmatrix} \mathbf{a}_j^T \\ \hline \mathbf{B} \end{bmatrix},$$

and can then write

$$\tilde{\mathbf{P}}_j \mathbf{X} = \begin{bmatrix} \mathbf{A}_j \\ \mathbf{B}_j \end{bmatrix}, \ \mathbf{A}_j \in \boldsymbol{R}^{(\mathrm{k\text{-}1}) \times \mathrm{n}}, \ \mathbf{B}_j \in \boldsymbol{R}^{(\mathrm{m\text{-}k+1}) \times \mathrm{n}}.$$

from which it follows that

$$\tilde{\mathbf{P}}_j \mathbf{X} = \begin{bmatrix} \mathbf{A}_j \\ \mathbf{B}_j \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_j \\ \mathbf{Y}_j \end{bmatrix} \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{\frac{1}{2}},$$

$$\mathbf{Q}_j = \mathbf{A}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-\frac{1}{2}}, \ \mathbf{Y}_j = \mathbf{B}_j \left(\mathbf{A}_j^T \mathbf{A}_j \right)^{-\frac{1}{2}}.$$

Note that,

$$\det\left(\mathbf{A}_j^T \mathbf{A}_j \right) = \det\left(\mathbf{A}^T \mathbf{A} - \mathbf{a}_j \mathbf{a}_j^T \right) = \left(1 - \left\| \mathbf{q}_j \right\|_2^2 \right) \det\left(\mathbf{A}^T \mathbf{A} \right).$$

Now let $\det\left(\mathbf{A}_j^T \mathbf{A}_j \right)$ be maximised when $p = j$. Then,

$$\det\left(\mathbf{A}_p^T \mathbf{A}_p \right) \geq \left(1 - \left\| \mathbf{q}_j \right\|_2^2 \right) \det\left(\mathbf{A}^T \mathbf{A} \right),$$

and, on summing over j we find that

$$k\det\left(\mathbf{A}_p^T\mathbf{A}_p\right)\geq (k-n)\det\left(\mathbf{A}^T\mathbf{A}\right). \qquad (11)$$

We can use this construction, starting with $\boldsymbol{X}$ and then deleting a row at the time whilst insuring that $\det\left(\mathbf{A}^T\mathbf{A}\right)$ is minimised at each step to construct

$$\boldsymbol{PX}=\begin{bmatrix}\boldsymbol{A}\\ \boldsymbol{B}\end{bmatrix},\ \ \boldsymbol{A}\in\boldsymbol{R}^{\mathrm{k}\times\mathrm{n}},\ \boldsymbol{B}\in\boldsymbol{R}^{(\mathrm{m}\text{-}\mathrm{k})\times\mathrm{n}},\ k>n\,\textbf{.}$$

From (11), it follows by induction that this construction satisfies

$$\det\left(\mathbf{A}^T\mathbf{A}\right)\geq\det\left(\mathbf{X}^T\mathbf{X}\right)\prod_{j=k+1}^{m}\frac{j-n}{j}=\frac{k!(m-n)!}{m!(k-n)!}\det\left(\mathbf{X}^T\mathbf{X}\right). \qquad \#$$

3 Discussion

We now compare the bounds given in corollary 1 and Theorem 3 which are the same for $n=1$. We have

$$\begin{aligned}\log\prod_{j=k+1}^{m}\left(\frac{j-n}{j}\right)&=\log\left(\frac{k+1-n}{k+1}\right)-\log\left(\frac{m+1-n}{m+1}\right)+\sum_{j=k+2}^{m+1}\log\left(\frac{j-n}{j}\right)\\&\geq\log\left(\frac{k+1-n}{k+1}\right)-\log\left(\frac{m+1-n}{m+1}\right)+\int_{k+1}^{m+1}\log\left(\frac{x-n}{x}\right)dx\\&=\log\frac{(m+1)(k+1-n)}{(k+1)(m+1-n)}+n\log\left(\frac{k+1-n}{m+1-n}\right)+(m+1)\log\left(1-\frac{n}{m+1}\right)-(k+1)\log\left(1-\frac{n}{k+1}\right)\\&=n\log\left(\frac{k+1-n}{m+1-n}\right)+m\log\left(1-\frac{n}{m+1}\right)-k\log\left(1-\frac{n}{k+1}\right).\end{aligned}$$

Thus, for $n\geq 2$

$$\log\prod_{j=k+1}^{m}\left(\frac{j-n}{j}\right)-n\log\left(\frac{k+1-n}{m+1-n}\right)\geq \underbrace{m\log\left(1-\frac{n}{m+1}\right)-k\log\left(1-\frac{n}{k+1}\right)}_{\geq 0}.$$

This demonstrates that the bound (10) given in Theorem 3 is superior to the bounds given by (8) in Corollary 1. This difference can be substantial when k is relatively small. For example, if m is large relative to n and $k=n$, then

$$\prod_{j=k+1}^{m}\left(\frac{j-n}{j}\right)/\left(\frac{k+1-n}{m+1-n}\right)^{n} \geq \left(1-\frac{n}{m+1}\right)^{m}/\left(1-\frac{n}{k+1}\right)^{k}$$
$$\approx \left(\frac{n+1}{e}\right)^{n}.$$

In order to compare the bounds on the singular values given by (6), it makes sense to consider the nth root of $\det \mathbf{A}^T\mathbf{A}$ as this is the square of the geometric mean of the singular values of $\mathbf{A}$. Given the construction, we find that the bound (8) given in Corollary 1 is similar to that given by (6). However, the bound given by (10) in Theorem 3, provides a substantially sharper estimate.

References

1. P.A. Businger and G.H. Golub, Linear least squares solution by Householder transformations, Numer. Math., 7 (1965), pp 269-276.

2. F.R. de Hoog and R.M.M. Mattheij, Subset Selection for Matrices, Linear Algebra and its Applications, 422 (2007), pp 349 – 359.

3. G.H. Golub and C.F. van Loan, Matrix Computations, John Hopkins University Press, Baltimore 1983.

4. C.L. Lawson and R.J. Hanson, Solving Least Squares Problems, Prentice Hall, Englewood Cliffs 1974.

5. Silvey SD, Optimal Design, Chapman and Hall, London, 1980.