

On "Cleaning" a Noisy Matrix by

Annuling its Smallest Singular Values

1

4/13/98

Robert W. Bass

Consider the problem of estimating an m -dimensional parameter-vector $p \in \mathbb{R}^m$ by regression on n observations $y \in \mathbb{R}^n$ where (with residual ^{vector} $u \in \mathbb{R}^n$)

$$y = Ap + u \quad (1)$$

and A is an $n \times m$ observation matrix.

More particularly, suppose that y, A, p are normally distributed random variables of mean values $\hat{y}, \hat{A}, \hat{p}$ and covariance matrices/tensors

$$C_y = E\{(y - \hat{y})(y - \hat{y})^T\} = E\{\tilde{y}\tilde{y}^T\}, \quad \tilde{y} \stackrel{d}{=} y - \hat{y}, \quad (2a)$$

$$C_A = E\{\tilde{A} \otimes \tilde{A}^T\}, \quad \tilde{A} \stackrel{d}{=} A - \hat{A}, \quad (2b)$$

$$C_p = E\{\tilde{p}\tilde{p}^T\}, \quad \tilde{p} \stackrel{d}{=} p - \hat{p}, \quad (2c)$$

where

$$E\{\tilde{y}\} = 0, \quad E\{\tilde{A}\} = 0, \quad E\{\tilde{p}\} = 0. \quad (3)$$

In order to make the problem tractable, we assume that (1) has been so "scaled" and/or "normalized" that the deviances \tilde{y} , \tilde{A} , $\tilde{\beta}$ are of "first order" where this means that "second order" deviances will be assumed to be negligible; specifically

$$\|\tilde{A}\tilde{\beta}\| \approx 0. \quad (4)$$

Then

$$\hat{y} = \hat{A}\hat{\beta} \quad (5)$$

and

$$v = \tilde{y} - \tilde{A}\hat{\beta} - \hat{A}\tilde{\beta} \quad (6)$$

so that

$$E\{v\} = E\{\tilde{y}\} - E\{\tilde{A}\}\hat{\beta} - \hat{A}E\{\tilde{\beta}\} = 0. \quad (7)$$

If we seek a minimal-variance estimator of β , we want to minimize $E\{\|\tilde{\beta}\|\} =$
 $= \text{trace}[E\{\tilde{\beta}\tilde{\beta}^T\}] = \text{trace}[C_p].$

Assuming this can be done, then the two main sources of error in the residual U are measurement errors \tilde{y} in y and noise \tilde{A} in A .

Assume that each element $\tilde{A} = (\tilde{a}_{i,j})$ is a zero-mean, ^{independent} white noise ^{process} and that all of these noises have the same "standard deviation" σ [i.e. same spectral intensity σ^2].

If A were noise-free, it is well known that one may minimize the expected value of the ^{squared} norm $\|U\|^2$ of the residual by using the estimator

$$\hat{\beta} = (A^T A)^{\dagger} A^T y \tag{8}$$

where M^{\dagger} denotes the pseudo-inverse of M .

If $\det(M) \neq 0$, then $M^{\dagger} = M^{-1}$.

4

Let

$$\sigma_i^2 \geq \sigma_{i+1}^2 \geq \dots \geq 0, \quad (i=1, \dots, m-1), \quad (9)$$

denote the eigenvalues of the symmetric

non-negative definite matrix $M = A^T A = M^T \geq 0$

arranged in descending order. Let $\{U^i\}$ denote

an orthonormalized set of m corresponding

[necessarily real]

eigenvectors ($M U^i = \sigma_i^2 U^i$) of M , and

define $V = (U^1, U^2, \dots, U^m)$. Then

$$V^T V = V V^T = I_m, \quad (10)$$

where $I_m = \text{diag}(1, 1, \dots, 1)$ denotes the $m \times m$

identity matrix. Moreover

$$V^T M V = S^2 = \text{diag}(\sigma_1^2, \dots, \sigma_m^2), \quad (11)$$

where for later use we define $S = \text{diag}(\sigma_1, \dots, \sigma_m)$

and take $\sigma_i = +\sqrt{\sigma_i^2} = |\sigma_i| \geq 0$.

Up to a rearrangement of columns

corresponding to multiple eigenvalues, V is unique.

5

For any [non-square] matrix $A_{n \times m}$ there can always be found ^{by the s.v.d. procedure} an $n \times m$ column-orthogonal matrix U , i.e.

$$U^T U = I_m, \quad (12)$$

and matrices S and V as in (10)-(11) such that

$$A = U S V^T. \quad (13)$$

Under the assumptions made, it is a remarkable fact that the SOLE EFFECT of the noise on A , as regards (8), is to add σ^2 to the squares of the singular values of \hat{A} !

Therefore one can completely remove the effect of noise on A by reducing all of the squared singular values of \hat{A} by the same [unknown] amount.

If the smallest $(J-m)$ squared singular

values are approximately equal

$$\sigma_k^2 \approx \sigma_{k+1}^2 \approx \dots \approx \sigma^2, \quad (k=J+1, \dots, m), \quad (14)$$

then their common value equals the variance of the noise on A . Accordingly one needs merely to use

$$\hat{\sigma}_{\text{OPT}, k}^2 = \sigma_k^2 - \sigma^2, \quad (k=1, 2, \dots, J), \quad (15a)$$

$$\sigma_{\text{OPT}, k}^2 = 0, \quad (k=J+1, \dots, m), \quad (15b)$$

and the estimator

$$\widehat{(A^T A)} = V \hat{S}^2 V^T, \quad (16)$$

in order to clean out the noise from A

preparatory to use of (8). [The noise on $(A^T y)$ is ^{order} second.]

Let

$$\hat{A} = U \hat{S} V^T \quad (17)$$

Then

$$\hat{A}^T \hat{A} = V \hat{S}^2 V^T. \quad (18)$$

Also

$$\begin{aligned} A^T A &= (\hat{A} + \tilde{A})^T (\hat{A} + \tilde{A}) = \\ &= \hat{A}^T \hat{A} + \hat{A}^T \tilde{A} + \tilde{A}^T \hat{A} + \tilde{A}^T \tilde{A}, \end{aligned} \quad (19)$$

Whence, using $E\{\tilde{A}\} = 0$,

$$E\{A^T A\} = \hat{A}^T \hat{A} + E\{\tilde{A}^T \tilde{A}\}. \tag{20}$$

Suppose that under the assumptions made

$$E\{\tilde{A}^T \tilde{A}\} = \sigma^2 I_m. \tag{21}$$

Then, because $I_m = V V^T$,

$$E\{A^T A\} = V \hat{S}^2 V^T + \sigma^2 I_m = V(\hat{S}^2 + \sigma^2 I_m) V^T, \tag{22a}$$

$$E\{V^T(A^T A)V\} = \hat{S}^2 + \sigma^2 I_m, \tag{22b}$$

where now the surprise is that the same V

which works for \hat{A} ^{in (17)} also, ^{as in (22a)} works for $(A^T A)$!

These considerations justify the widely-known heuristic procedure for "cleaning" A by means of (15), whenever (14) holds, provided only that (21) can be proved.

This seems to require use of Kronecker [tensor] products, as explained in pages ①-⑤ of the following Appendix.

Conclusions

Regardless of the changes in size of the larger, ^{squared} singular values (which can cover a spread of many orders of magnitude) the effect of noise on A always is most pronounced in the smallest, ^{squared} singular value. If this one is significantly smaller than the next-to-smallest, it is reasonable to suppose that it represents the intensity of the noise on all of the larger ones, and should be subtracted from all of them, including itself. Especially if there is a cluster of small, ^{squared} singular values of approximately equal magnitudes, it is reasonable to suppose that their common value is the result of noise, and to subtract this value from the entire set.

[APPENDIX]

$$A = (a^1, a^2, \dots, a^m), \quad a^i \in \mathbb{R}^n, \quad (i=1, 2, \dots, m) \quad \textcircled{1}$$

$$A = (a_i^j \mid (i=1, 2, \dots, n), (j=1, 2, \dots, m)) = A_{n \times m}$$

$$\underline{a} = \text{vec}(A) = \begin{pmatrix} a^1 \\ a^2 \\ \vdots \\ a^m \end{pmatrix} \in \mathbb{R}^{mn}$$

$$A = \text{mat}(\underline{a}), \quad \underline{a} \in \mathbb{R}^{mn}$$

$$\text{kron}(M_1, M_2) \stackrel{d}{=} ((M_1)_{ij} M_2) \stackrel{d}{=} M_1 \otimes M_2.$$

$$A = (a^1, \dots, a^m), \quad B = (b^1, \dots, b^p), \quad b^i \in \mathbb{R}^m$$

$$\Rightarrow A \cdot B = (Ab^1, \dots, Ab^p) = \text{mat} \begin{pmatrix} Ab^1 \\ \vdots \\ Ab^p \end{pmatrix} =$$

$$= \text{mat}(\text{kron}(I_p, A) \text{vec}(B))$$

$$= \text{mat} \left(\begin{bmatrix} A, \phi, \dots, \phi \\ \phi, A, & & \\ \vdots & \ddots & \\ \phi, & & A \end{bmatrix} \begin{pmatrix} b^1 \\ b^2 \\ \vdots \\ b^p \end{pmatrix} \right)$$

(2)

$$B.A = (Ba^1, \dots, Ba^m) = \text{mat} \begin{pmatrix} Ba^1 \\ \vdots \\ Ba^m \end{pmatrix} =$$

$$= \text{mat} \begin{pmatrix} a_1^1 b^1 + \dots + a_n^1 b^p \\ \vdots \\ a_1^m b^1 + \dots + a_n^m b^p \end{pmatrix} =$$

$$= \text{mat} \left(\begin{bmatrix} a_1^1 I_n & a_2^1 I_n & \dots & a_n^1 I_n \\ \vdots & \vdots & \ddots & \vdots \\ a_1^m I_n & \dots & \dots & a_n^m I_n \end{bmatrix} \begin{pmatrix} b^1 \\ \vdots \\ b^p \end{pmatrix} \right)$$

$$= \text{mat} \left([\text{kron}(A^T, I_n)] \text{vec}(B) \right).$$

if $P = P^T$ & A are $n \times n$,

$$\text{Hence, } \text{vec}(PA + A^T P) =$$

$$= [A^T \otimes I_n + I_n \otimes A^T] \cdot \text{vec}(P)$$

$$\text{So } PA + A^T P = -I_n \iff \det(A^T \otimes I_n + I_n \otimes A^T) \neq 0$$

$$\& P = \text{mat}(\mathbb{P}), \mathbb{P} = -(A^T \otimes I_n + I_n \otimes A^T)^{-1} \text{vec}(I_n)$$

$$\text{where } I_n = \text{diag}(1, 1, \dots, 1) = (e^1, e^2, \dots, e^n) \& \text{vec}(I_n) = \begin{pmatrix} e^1 \\ e^2 \\ \vdots \\ e^n \end{pmatrix}.$$

(3)

By definition

Example, $A_{n \times n}$ is Hurwitz $\Leftrightarrow \exists \gamma \geq 1 \ \& \ \alpha > 0$

$\exists \ \|e^{At}\| \leq \gamma e^{-\alpha t}, \ \forall t \geq 0.$

If A is Hurwitz, then

$$\|e^{A^T t} e^{At}\| \leq \gamma^2 e^{-2\alpha t}, \ \forall t \geq 0,$$

$$\Rightarrow \exists \ P = \int_0^{+\infty} e^{A^T \tau} e^{A\tau} d\tau = P^T \ \&$$

$$\langle x, Px \rangle = x^T Px = \int_0^{+\infty} \|e^{A\tau} x\|^2 d\tau > 0 \ \forall x \neq 0$$

& so $P = P^T > 0$ is positive definite.

Also P satisfies $PA + A^T P + I_n = 0$ because

$$I_n + PA + A^T P = I_n + \int_0^{+\infty} (A^T e^{A^T \tau} e^{A\tau} + e^{A^T \tau} e^{A\tau} A) d\tau =$$

$$= I_n + \lim_{T \rightarrow +\infty} \int_0^T \frac{d}{d\tau} (e^{A^T \tau} e^{A\tau}) d\tau =$$

$$= I_n + \lim_{T \rightarrow +\infty} \left[e^{A^T T} e^{AT} - e^{A^T \cdot 0} \cdot e^{A \cdot 0} \right] =$$

$$= I_n + 0 \cdot I_n - I_n = 0, \text{ because } e^{A \cdot 0} = I_n.$$

Now suppose $A = \text{mat}(\underline{a})$, $\underline{a} \in \mathbb{R}^{mn}$,

$$A = (a^1, a^2, \dots, a^m), \quad a^i \in \mathbb{R}^n, \quad (i=1, 2, \dots, m).$$

Also suppose each element $a_{i,j}$ of A is a Gaussian random variable, with mean $\hat{a}_{i,j}$ & standard deviation σ , such that

$$E\{a_{i,j}\} = \hat{a}_{i,j}, \quad E\{(a_{i,j} - \hat{a}_{i,j})^2\} = \sigma^2$$

Also suppose the $\tilde{a}_{i,j} = a_{i,j} - \hat{a}_{i,j}$ are UNCORRELATED so $E\{\tilde{a}_{i,j} \tilde{a}_{k,l}\} = \sigma^2 \delta_{ij,kl}$

Then $A = \hat{A} + \tilde{A}$ where $\hat{A} = E\{A\}$ & $\tilde{A} = A - \hat{A}$.

Obviously $E\{\tilde{A}\} = 0$. What is $E\{\tilde{A}^T \tilde{A}\}$?

Note that $(\tilde{A})^T \tilde{A} =$

$$= \text{mat}([I_m \otimes \tilde{A}^T] \text{vec}(\tilde{A})) =$$

$$\text{mat} \left(\begin{bmatrix} \tilde{A}^T & \phi & \dots & \phi \\ \phi & \tilde{A}^T & \dots & \phi \\ \vdots & & & \\ \phi & \dots & & \tilde{A}^T \end{bmatrix} \begin{pmatrix} \tilde{a}^1 \\ \tilde{a}^2 \\ \vdots \\ \tilde{a}^m \end{pmatrix} \right)$$

$$\text{Now } \tilde{A}^T \tilde{a}^n = \begin{bmatrix} (\tilde{a}^1)^T \\ \vdots \\ (\tilde{a}^m)^T \end{bmatrix} \tilde{a}^n = (\langle \tilde{a}^i, \tilde{a}^n \rangle), \&$$

$$E\{\tilde{A}^T \tilde{a}^n\} = (E\{\langle \tilde{a}^i, \tilde{a}^n \rangle\}) = (\sigma^2 \delta_{in}), \text{ so}$$

5

if $I_m = (e^1, e^2, \dots, e^m)$, $e^i \in \mathbb{R}^m$, then

$$\mathcal{E}\{\tilde{A}^T \tilde{a}^r\} = e^r. \quad \text{Hence } (\tilde{A})^T \tilde{A} = \text{mat}(I_m \otimes (\tilde{A})^T \text{vec}(\tilde{A}))$$

implies that

$$\mathcal{E}\{(\tilde{A})^T \tilde{A}\} = \text{mat} \begin{pmatrix} \sigma^2 e^1 \\ \sigma^2 e^2 \\ \vdots \\ \sigma^2 e^m \end{pmatrix} = \sigma^2 I_m, \text{ as}$$

originally claimed.