Asymptotic behaviour in linear least squares problems

M.R.Osborne *

Abstract

The asymptotic behaviour of a class of least squares problems when subjected to structured perturbations is considered. It is permitted that the number of rows (observations) in the design matrix can be unbounded while the number of degrees of freedom (variables) is fixed. It is shown that for certain classes of random data the solution sensitivity depends asymptotically on the condition number of the design matrix rather than on its square which is the generic result for inconsistent systems when the norm of the residual is not small. Extension of these results to the case where the perturbations are due to rounding errors is considered.

1 Introduction

The linear least squares problem has the general form

$$\min_{\mathbf{x}} \mathbf{r}^T\mathbf{r};\ \mathbf{r} = A\mathbf{x} - \mathbf{b}. \tag{1}$$

where the design matrix $A : R^p \to R^n$, the residual and observation vectors $\mathbf{r}$, $\mathbf{b} \in \mathbf{R}^n$, and the vector of model parameters $\mathbf{x} \in R^p$. Typically p will be fixed corresponding to a known model, while n will usually be assumed "large enough". Limiting processes will assume that p is fixed and $n \to \infty$.

This problem is a simple optimization problem subject to equality constraints. The necessary conditions for a minimum give

$$0 = \nabla_x \mathbf{r}^T\mathbf{r} = \mathbf{2r}^T A. \tag{2}$$

*Mathematical Sciences Institute, Australian National University, ACT 0200, AUSTRALIA. mailto:Mike.Osborne@anu.edu.au

This paper is based on a presentation given to the meeting honouring the lives of Gene Golub and Ron Mitchell held at the Australian National University on February 28-29, 2008. It is dedicated to the memory of these two good friends.

Substituting for $\mathbf{r}$ from (1) gives the *normal equations*

$$A^T A\mathbf{x} = A^T\mathbf{b}. \tag{3}$$

This system defines uniquely both the least squares estimator $\mathbf{x}^{(n)}$ and the corresponding residual vector $\mathbf{r}^{(n)}$ providing the design matrix A has full column rank p, and this condition is assumed.

An important modelling context which generates linear least squares problems is the following. Assume noisy observations are made on a system at a sequence of configurations labelled by a reference variable t which could be time. Let these be summarised by

$$b_i = y(t_i) + \varepsilon_i,\ i = 1, 2, \cdots, n, \tag{4}$$

where $y(t)$ is the error free signal (true model) which is assumed to be expressible in parametric form as

$$y(t) = \sum_{i=1}^{p} x_i^* \phi_i(t), \tag{5}$$

the $x_i^*,\ i = 1, 2, \cdots, p$, are the (hypothesised) true parameter values, the $\phi_i(t),\ i = 1, 2, \cdots, p$ are basis functions specifying the model class, and the ε_i are random variables summarising the noise in the observations. A standard assumption would be that the ε_i are independent and normally distributed with mean 0, and standard deviation γ ($\varepsilon \backsim N(0, \gamma^2 I)$). In this case the solution of equation (3) is the maximum likelihood estimate and has important minimum variance properties. However, the solution $\mathbf{x}$ provides consistent parameter estimates for significantly more general families of distributions. The property of (strong) consistency ensures that

$$\mathbf{x}^{(n)} \to \mathbf{x}^*,\ \text{almost surely}, n \to \infty.$$

It requires conditions on the choice of $\{t_i, i = 1, \cdots, n\}$ which are satisfied by the sampling properties required below. For this model

$$A_{ij} = \phi_j(t_i),\ j = 1, 2, \cdots, p,\ i = 1, 2, \cdots, n.$$

Remark 1 *Note that in this context* $\frac{1}{n}\|\mathbf{r}\|^2$ *estimates the variance* γ^2*. This means that the residual vector can be expected to be bounded away from zero in norm for* $n > p$*. It follows that without further information this class of problems would be expected to fall into the class in which a* $(\text{cond}\, A)^2$ *error estimate is expected.*

It is important to know how the estimate $\mathbf{x}^{(n)}$ of $\mathbf{x}^*$ given by (3) behaves as the number of observations increases without limit because this permits statements to be made about the rates of convergence implied by the consistency of the estimates. This is not just a theoretical point because it informs on how much data needs to be collected and so directly relates to the practicality of the measurement exercise. It is also important to know how the computational algorithm chosen to solve (1) will behave on large data sets. In this connection, the first point to make is that a *systematic* process capable of automation is required to generate the values of the reference variable t_i and record the observations b_i associated with large data sets if an asymptotic analysis is to be possible. The nature of this recording process must depend on the nature of the system being observed. It is assumed that the system has the property that after a finite horizon, assumed to correspond to values $0 \leq t \leq 1$ for the labelling variable, no further information on model structure is available. One case corresponds to signals decaying to zero. However, the case of a finite observation window dictated by external factors is also included. Such systems may be required to be controlled, and may have quite complicated stability properties.

The refinement process used to increase n is one in which independent trials are performed to obtain data for sequences of points $\left\{t_i^{(n)},\ i = 1, 2, \cdots n\right\}$ for an increasing sequence of values of n. The use of the descriptor systematic is taken to mean that there is a limiting process such that

$$\frac{1}{n}\sum_{i=1}^{n} f(t_i^{(n)}) \rightarrow \int_0^1 f(t)dw(t),\ n \rightarrow \infty, \tag{6}$$

holds for all sufficiently smooth $f(t)$ ($f(t) \in C\,[0,1]$ for example) where $w(t)$ is a weight function characteristic of the sampling regime. The left hand side in (6) can be interpreted as a simple quadrature formula. For example, $w(t) = t$ in the two cases:

1. The t_i are equispaced. The corresponding quadrature error for smooth enough $f(t)$ is strictly $O(1/n)$.

2. The t_i are uniformly distributed in $[0,1]$. The corresponding quadrature error is asymptotically normally distributed with variance $O\,(1/n)$.

Such samplings are called *regular* to stress that there is a sense in which the quadrature error is $o(1)$, $n \rightarrow \infty$. The associated convergence mode is denoted "r.e.". For example, $\underset{n\rightarrow\infty}{\overset{r.e.}{\rightarrow}}$. The particular sense appropriate is implied.

Now assume that the design matrix in (1) is constructed for each n using a regular sampling procedure. Then the regularity condition gives

$$\frac{1}{n}A_{*i}^T A_{*j} = \frac{1}{n}\sum_{k=1}^{n}\phi_i(t_k^{(n)})\phi_j(t_k^{(n)}) \\ \overset{r.e.}{\underset{n\to\infty}{\to}} \int_0^1 \phi_i(t)\phi_j(t)dw(t) = G_{ij}. \tag{7}$$

This states that the normal matrix in (3) scaled by $\frac{1}{n}$ approaches the Gram matrix $G : R^p \to R^p$ of the set $\Phi = \{\phi_j(t), j = 1, 2, \cdots, p\}$ relative to the weight $w(t)$ with an error which is $o(1)$, $n \to \infty$. This rate is assumed fast enough to ensure $\operatorname{cond} A \to \operatorname{cond} G$, $n \to \infty$. G is nonsingular and positive definite by assumption. It need not be well conditioned. For example, values of spectral condition numbers for Hilbert matrices corresponding to the case when elements of Φ are monomials and the t_i are equispaced are given in [4] for $2 \le n \le 16$. One consequence of the above discussion is that there is no real restriction in assuming that the subordinate matrix norm relative to the euclidean norm of the design matrix satisfies $\|A\| = \sqrt{n}$. This amounts to a rescaling of the design A by a quantity which is asymptotically constant.

The next section treats some consequences of perturbing the data of equation (1). The classic inequality of Golub and Wilkinson is derived and certain asymptotic properties for large n are explored. In particular, the influence of the stochastic components in the data vector $\mathbf{b}$ is considered. This adds a somewhat different perspective to the usual worst case scenarios because here the law of large numbers [6] in the (extended) form

$$\frac{1}{n}\sum_{i=1}^{n} X_{ni}\varepsilon_{ni} \overset{a.s.}{\underset{n\to\infty}{\to}} 0 \tag{8}$$

is available when the ε_{ni} are independent and of bounded variance for all n, and the constants X_{ni} are bounded. This result permits the influence of the "bad term" involving the square of the condition number of A to be ignored in certain circumstances. The resulting perturbation behaviour then becomes similar to that for consistent linear systems.

2 Perturbation of least squares problems

We consider the generic perturbed least squares problem (1) with data

$$\mathbf{r} = (A + \tau E)\,\mathbf{x} - (\mathbf{b} + \tau \mathbf{z})$$

where perturbations E, $\mathbf{z}$ are fixed in the sense that they result from a well defined rule for each n. The perturbation E is assumed to be independent of any observational error. It is assumed that τ is a small parameter. The component-wise scale of the perturbations is fixed by requiring

$$\max_{i,j} |E_{ij}| = \eta \leq 1, \ \|\mathbf{z}\|_\infty \leq 1. \tag{9}$$

It follows that $\frac{1}{n}E^T A$ has uniformly bounded elements. We have

$$\max_{1\leq i,j\leq p} \frac{1}{n}|(E^T A)_{ij}| \leq \frac{1}{n}\max_{i,j}\sum_{s=1}^{n}|E_{si}||A_{sj}| \leq \eta \max_{i,j}|A_{ij}| \leq \eta \max_i \|\phi_i\|_\infty. \tag{10}$$

It is assumed that τ is small enough for both A and $A+\tau E$ to have their full rank p. The necessary conditions for the perturbed and unperturbed least squares problems are

$$(A+\tau E)^T \widehat{\mathbf{r}} = 0, \ A^T \mathbf{r}^{(n)} = 0$$

where the $\widehat{\ }$ indicates the solution of the perturbed problem. Subtracting gives

$$(A+\tau E)^T \left(\widehat{\mathbf{r}} - \mathbf{r}^{(n)}\right) + \tau E^T \mathbf{r}^{(n)} = 0,$$

and substituting for the residual vectors gives the basic relation

$$(A+\tau E)^T (A+\tau E)\left(\widehat{\mathbf{x}} - \mathbf{x}^{(n)}\right) = \tau\left\{(A+\tau E)^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - E^T\mathbf{r}^{(n)}\right\}. \tag{11}$$

For small enough τ and each fixed n this gives

$$\begin{aligned}\widehat{\mathbf{x}} - \mathbf{x}^{(n)} &= \tau\left\{\left(A^T A\right)^{-1}\left(A^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - E^T\mathbf{r}^{(n)}\right)\right\} + O\left(\tau^2\right), \\ &= \tau\left\{\begin{array}{c}\left(\frac{1}{\sqrt{n}}U\right)^{-1}\frac{1}{\sqrt{n}}Q_1^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right) \\ -\left(\frac{1}{n}A^T A\right)^{-1}\frac{1}{n}E^T\mathbf{r}^{(n)}\end{array}\right\} + O\left(\tau^2\right),\end{aligned} \tag{12}$$

where A possesses the orthogonal Q times upper triangular U factorization

$$A = Q\begin{bmatrix} U \\ 0 \end{bmatrix} = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix}\begin{bmatrix} U \\ 0 \end{bmatrix} = Q_1 U,$$

and Q_1 corresponds to the first p columns of Q. There are two ways of looking at this relation. The first considers n fixed and worries about the

size of $\mathrm{cond}\,(A) = \frac{\sigma_1}{\sigma_p}$, the ratio of the largest to smallest singular values of A. This leads to the basic inequality

$$\left\|\widehat{\mathbf{x}} - \mathbf{x}^{(n)}\right\| \leq \tau \left\{ \frac{\mathrm{cond}\,(A)}{\sqrt{n}} \left\|\mathbf{z} - E\mathbf{x}^{(n)}\right\| + \frac{\mathrm{cond}\,(A)^2}{n} \left\|E^T\mathbf{r}^{(n)}\right\| \right\} + O\left(\tau^2\right), \tag{13}$$

where the assumption that $\|A\| = \sqrt{n} = \sigma_1$ has been used. The original form of this result is due to [3]. Equation (13) reveals the possible dominance of the term $\mathrm{cond}\,(A)^2$. This is likely if $\frac{1}{n}\left\|E^T\mathbf{r}^{(n)}\right\|$ is not small. The importance of the inequality (13) is that it is a generic result highlighting what is best possible. For this reason computational algorithms in which the error takes this form are said to have *optimal error structure*. Development of such optimal algorithms based on the use of orthogonal transformations goes back to [5], [2], and [1]. It follows from (12) that

$$\begin{aligned} \widehat{\mathbf{r}} - \mathbf{r}^{(n)} &= -\tau \left\{ (I-P)\,\mathbf{z} + PE\mathbf{x}^{(n)} + A\left(A^TA\right)^{-1}E^T\mathbf{r}^{(n)} \right\} + O\left(\tau^2\right), \\ &= -\tau \left\{ (I-P)\,\mathbf{z} + PE\mathbf{x}^{(n)} + Q_1U^{-1}E^T\mathbf{r}^{(n)} \right\} + O\left(\tau^2\right) \end{aligned} \tag{14}$$

where P is the orthogonal projection $A\left(A^TA\right)^{-1}A^T$ onto the range of A. Thus the result of the perturbation is a change of magnitude $O\left(\mathrm{cond}\,(A)\right)$ in the residual showing that a more satisfactory result is possible if the computed residual is the required quantity.

However, there is an alternative way of considering this result which is important when n is large and $\boldsymbol{\varepsilon}$ is a random vector. The Gram matrix G (7) is used to write a limiting form of (12) as $n \overset{r.e.}{\to} \infty$. Contributions from quadrature error terms in this approximation have been ignored (Lemma 3 shows they contribute at most $\tau\,(o(1))$ for large n given regular sampling), and $G^{1/2}$ is written for the large n approximation to $\frac{1}{\sqrt{n}}U$ in (12).

$$\widehat{\mathbf{x}} - \mathbf{x}^{(n)} = \tau \left\{ G^{-1/2}\frac{1}{\sqrt{n}}Q_1^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right) - G^{-1}\frac{1}{n}E^T\mathbf{r}^{(n)} \right\} + O\left(\tau\left(o(1)\right), \tau^2\right). \tag{15}$$

We have the following bounds for the interesting terms in this equation.

Lemma 2

$$\frac{1}{\sqrt{n}}\left\|Q_1^T\left(\mathbf{z} - E\mathbf{x}^{(n)}\right)\right\| \leq \left\|\mathbf{z} - E\mathbf{x}^{(n)}\right\|_\infty,$$

$$\frac{1}{n}\left\|E^T\mathbf{r}^{(n)}\right\| \leq \sqrt{\frac{p}{n}}\eta\left\|\mathbf{r}^{(n)}\right\|_\infty.$$

Proof. The first part applies the standard inequality relating the 2 and ∞ norms. The second part follows in similar fashion from the inequality

$$\|E\| \leq \sqrt{np}\eta, \tag{16}$$

where the right hand side is a simple bound for the Frobenius norm of E. ■

Also it is important when fixing the order of dependence on τ that the n dependence of the $O(\tau)$ terms is appropriately bounded as $n \overset{r.e.}{\to} \infty$. The key is the following result.

Lemma 3

$$\frac{1}{n}(A+\tau E)^T (A+\tau E) \underset{n\to\infty}{\overset{r.e.}{\to}} G + O(\tau).$$

It follows that the normal matrix associated with the perturbed least squares problem has a suitably bounded inverse under regular sampling.

Proof.

$$\begin{aligned}\frac{1}{n}(A+\tau E)^T (A+\tau E) = \frac{1}{n}U^T \left\{ I + \tau \left\{ Q_1^T E U^{-1} + U^{-T} E^T Q_1 \right\} \right. \\ \left. + \tau^2 U^{-T} E^T E U^{-1} \right\} U.\end{aligned} \tag{17}$$

To show that the terms multiplying both τ and τ^2 in this expression are $O(1)$, $n \overset{r.e.}{\to} \infty$, requires a bound for $\|EU^{-1}\|$ valid for large n. Note $\|EU^{-1}\| \geq \left\|Q_1^T E U^{-1}\right\|$. The required bound can be constructed as follows:

$$\begin{aligned}\left\|U^{-T} E^T E U^{-1}\right\| &= \sup_{\mathbf{v}} \frac{\mathbf{v}^T U^{-T} E^T E U^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}}, \\ &= \sup_{\mathbf{w}} \frac{\mathbf{w}^T E^T E \mathbf{w}}{\mathbf{w}^T U^T U \mathbf{w}}, \\ &\leq \frac{\|E\|^2}{n\sigma_{min}\left\{\frac{1}{n}A^T A\right\}}, \\ &\leq \frac{p\eta^2}{\sigma_{min}\{G\}} + o(1),\ n \overset{r.e.}{\to} \infty,\end{aligned}$$

where the estimate of $\|E\|$ given in the previous Lemma has been used. Thus

$$\left\| \frac{1}{n}(A+\tau E)^T (A+\tau E) - G \right\| \leq \tau \left\{ 3\|G\|^{1/2} \sqrt{p\,\mathrm{cond}(G)} + o(1) \right\},\ n \overset{r.e.}{\to} \infty.$$

The last step uses $\tau^2 \|EU^{-1}\|^2 \leq \tau \|EU^{-1}\|$ when $\tau \|EU^{-1}\| \leq 1$. ■

Remark 4 *This result has the consequence that all "Big O" terms in the basic relation (15) have the orders claimed as* $n \overset{r.e.}{\to} \infty$. *More can be said. We have*

$$E^T \mathbf{r}^{(n)} = E^T \left\{ A \left(\mathbf{x}^{(n)} - \mathbf{x}^* \right) - \boldsymbol{\varepsilon} \right\}.$$

It is necessary to use both consistency and the law of large numbers to estimate these contributions. Consistency gives

$$\frac{1}{n} E^T A \left(\mathbf{x}^{(n)} - \mathbf{x}^* \right) \to 0, \ \textit{almost surely},$$

using the boundedness of the elements of the $p \times p$ *matrix* $\frac{1}{n} E^T A$ *(10); and*

$$\frac{1}{n} E^T \boldsymbol{\varepsilon} \to 0, \ \textit{almost surely},$$

by the law of large numbers (8). It follows that there is a sense in which the term in $G^{-1/2}$ *dominates in (15) for large* n.

3 What about rounding error?

This section considers the possibility of applying the previous results to the important class of problems in which E, $\mathbf{z}$, τ are determined by the computational procedure and the characteristics of floating point arithmetic. We consider the Golub orthogonal factorization algorithm based on Householder transformations [2]. An appropriate error analysis from [4], Theorem 20.3, is summarised here.

"Let $A \in R^{n \times p}$ $(n \geq p)$ have full rank. The computed solution $\widehat{\mathbf{x}}$ of (1) is the exact least squares solution of the perturbed problem

$$\min_{\mathbf{x}} \| \mathbf{b} + \tau \mathbf{z} - (A + \tau E) \, \mathbf{x} \|,$$

where the perturbations satisfy the component wise bounds

$$|\tau E| \leq np \gamma_{cm} G |A|, \quad |\tau \mathbf{z}| \leq np \gamma_{cm} G |\mathbf{b}|,$$

where $\| G | |_F = 1$ and it is assumed that the constant γ_{cm} satisfies $np\gamma_{cm} < \frac{1}{2}$."

The perturbation scale τ is determined by the requirement that the component-wise scaling conditions (9) are satisfied. It is clear that it is related to γ_{cm} and so to bounds for accumulated round-off error. This presents little practical difficulty until it comes to consideration of asymptotic behaviour for very large n. Practical experience would seem to indicate that the asymptotics do work for large but practical values of n.

However, rounding errors give rise to another class of questions as the dependence of $\mathbf{b}$ on $\boldsymbol{\varepsilon}$ means that stochastic and rounding errors must be coupled to some extent. Clearly this affects $\mathbf{z}$. But it also affects E. This is expressed in [4] as

$$\tau E = \Delta A + Q_1 \Delta R$$

where ΔA comes from the computed factorization of A and is independent of $\mathbf{b}$, Q is an orthogonal matrix, and ΔR takes account of the contribution from the back substitution and so involves some coupling between rounding and stochastic errors. It follows that the argument of the previous section can be applied to show that the contribution from the term ΔA can be made small. However, it also means that the condition of independence implicit in the consistency and law of large numbers estimates no longer applies so that these tools do not suffice to show the contribution from ΔR is similarly small.

The importance of the law of large numbers is that it gives a method for quantifying cancellation in statistical calculations. There seems to be abundant evidence that cancellation to reduce the impact of rounding error accumulation is also a significant aid in large scale computations. It would be distinctly useful to have some form of analogue of the law of large numbers as a tool in these circumstances. Note that the law of large numbers does not preclude exceptional cases. It just guarantees they are very small in number. A computational analogue would similarly have to permit a small number of worst case scenarios.

4 Acknowledgement

The author is appreciative of the interest shown by the referees and by Nick Higham. Their comments have led to important improvements.

References

[1] Å. Björck, *Solving linear least squares problems by Gram-Schmidt orthogonalization*, BIT, 7 (1967), pp. 1–22.

[2] G. H. Golub, *Numerical methods for solving least squares problems*, Num. Math., 7 (1965), pp. 206–216.

[3] G. H. Golub and J. H. Wilkinson, *Iterative refinement of least squares solutions*, Num. Math., 9 (1966), pp. 189–198.

[4] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, 2002. Second Edition.

[5] A. S. Householder, *Unitary triangularization of a nonsymmetric matrix*, J. ACM, 6 (1958), pp. 339–342.

[6] K. S. Sen and J. M. Singer, *Large Sample Methods in Statistics*, Chapman and Hall, 1993.