DP2007/06

Conditioning and Hessians in analytical and numerical optimisation: Some illustrations

Christie Smith

April 2007

JEL classification: C61, C63

www.rbnz.govt.nz/research/discusspapers/

Discussion Paper Series

ISSN 1177-7567

DP2007/06

Conditioning and Hessians in analytical and numerical optimisation: Some illustrations*

Christie Smith†

Abstract

This note illustrates the connections between the Hessians of numerical optimisation problems, variance-covariance matrices for parameter vectors, and the influence that data mismeasurement may have on parameter estimates. Condition numbers provide a central guide to the sensitivity of common numerical problems to data mismeasurement. Examples are provided that clarify their importance. Two simple prescriptions arise from this analysis. First, data must be of an 'appropriate' scale. In some cases this means that the data need similar means *and* similar variances. Second, in numerical algorithms it is desirable to ascertain the condition number of the Hessian implied by the initial parameter values used for numerical optimisation algorithms. Condition numbers are easy to compute and indicate whether the updates from an initial starting value are likely to be poor.

* The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Reserve Bank of New Zealand. I would like to thank Andrew Coleman and Rishab Sethi for helpful comments on this note.

† Corresponding author: Christie Smith, Economics Department, Reserve Bank of New Zealand, 2 The Terrace, PO Box 2498, Wellington, New Zealand. E-mail address: Christie.Smith@rbnz.govt.nz.

ISSN 1177-7567

1 Introduction

This note illustrates the connections between the Hessians of numerical optimisation problems, variance-covariance matrices for parameter vectors, and the influence that data mismeasurement may have on parameter estimates. Section 2 briefly notes the connection between linear regression and the classic linear problem. Section 3 then illustrates how solutions to these problems respond to data mismeasurement and provides an upper bound on the elasticity of the solution to the mismeasurement. Section 5 then illustrates that the linear problems may also feature in numerical optimisation techniques, and the condition of such problems is particularly important.

Two simple prescriptions arise from this analysis. First, data must be of an 'appropriate' scale. In some cases this means that the data need similar means *and* similar variances, and ideally the dimensions of the data will be orthogonal. Second, in numerical algorithms it is desirable to ascertain the condition number of the Hessian implied by the initial parameter values used for numerical optimisation algorithms. Condition numbers are easy to compute and indicate whether the updates from an initial starting value are likely to be poor.

2 The linear problem in linear regression

Consider the following system of linear equations:

$$b = Ax \tag{1}$$

The desired solution from this linear equation is an estimate of x that satisfies the relationship between b, A, and x, where b and A are known. Theoretically one can simply use the (possibly generalized) inverse of A to obtain the solution x, but in practice, this inversion may be computationally challenging. Consider this problem in the context of linear regressions. For linear regression the linear problem (1) is associated with the first order conditions used to obtain the parameter estimates.

Denote the linear *regression* function:

$$y = W\gamma + \epsilon \tag{2}$$

where y is a $T \times 1$ vector of the dependent variable; W is a $T \times n$ matrix of independent variables; γ is the $n \times 1$ vector of parameters and ϵ is a $T \times 1$ vector of errors. The standard least squares approach to solving this regression problem is to estimate γ by minimising the sum of squared errors:[1]

$$\epsilon'\epsilon = (y - W\gamma)'(y - W\gamma)$$

which yields the following first order conditions with respect to γ:

$$W'y - W'W\gamma = 0$$

Let $b = W'y$, $A = (W'W)$ and $x = \gamma$. Then it is obvious that the first order conditions are linear equations as in equation (1).

3 Perturbations in b, A, or x

In the linear problem described in (1) one can assess the sensitivity of results to perturbations – or mismeasurements – in b, A, or x, or some combination thereof. Judd (1999,67) explores the case where there is mismeasurement only in b. In the linear regression case this corresponds to mismeasurement only in the regressand. One could equally envisage a case where the mismeasurement occurs in the regressor matrix A. Data mismeasurement is particularly important in the context of computational economics. Since computers can only store a finite number of digits when representing a number there are limitations on computers' ability to conduct operations accurately. Interestingly, the properties of A provide a bound that determines the effect of the perturbations in each case.

3.1 Perturbations in b

Suppose that there are errors in b but no errors in A. In the linear regression context this means that there can be errors in the dependent variable y but

[1] It should be noted that there are other algorithms, such as using the QR decomposition, which have better numerical properties for computing least squares parameter estimates.

not in the regressors W. Let $\tilde{x}$ be the solution of the system when b has been subjected to an error δb, ie

$$b + \delta b = A\tilde{x}$$

Let the error in $\tilde{x}$ be defined as $\delta x \equiv (\tilde{x} - x)$. Then by the linearity of the equation

$$\delta b = A\left(\tilde{x} - x\right) = A\delta x \tag{3}$$

Implying that

$$A^{-1}\delta b = \delta x = (\tilde{x} - x) \tag{4}$$

The relative condition number (Trefethen and Bau, 1997) is:

$$\frac{\|\delta x\|}{\|x\|} \div \frac{\|\delta b\|}{\|b\|} \tag{5}$$

where $\|\cdot\|$ denotes a vector norm (and the corresponding induced matrix norm).[2] Essentially, this is an elasticity. The denominator in this fraction corresponds to a kind of 'percent error' in the data, while the numerator corresponds to the 'percent error' in the parameter vector that we are trying to estimate.[3] Ideally, this elasticity will be small, so that data errors (the denominator) do not generate large errors in the parameter estimate (the numerator).[4] Recall that we are trying to minimise the error in the solution $\widetilde{x}$, where x is the vector of parameters.

The above elasticity, the relative condition number, is related to the condition number of the matrix $A = W'W$. In other words, we need to worry about the properties of A in assessing the vulnerability of our solution to data

[2] A vector norm $\|\cdot\|$ is a functional on an m-dimensional complex space, $\|\cdot\| : \mathbb{C}^m \to \mathbb{R}$, that has the following properties, $\forall$ vectors x and y in the space: i) $\|x\| \geq 0$, with equality $\iff$ $x = 0$; ii) $\|\alpha x\| = |\alpha| \cdot \|x\|$ $\forall$ scalar α; iii) $\|x\| + \|y\| \geq \|x + y\|$.The inner product is a norm; ie if x is a vector of dimension k then $\|x\| = \left(\sum_{i=1}^{k} x_i^2\right)^{1/2}$ is a norm. An *induced matrix norm* $\|A\|_{(m,n)} = \sup_{\{x \in \mathbb{C}^n, x \neq 0\}} \frac{\|Ax\|_{(m)}}{\|x\|_{(n)}}$, where $\|\cdot\|_{(m)}$ and $\|\cdot\|_{(n)}$ are norms on the range and domain of $A \in \mathbb{C}^{m \times n}$. Equivalently $\|A\|_{(m,n)} = \sup_{\{x \in \mathbb{C}^n, \|x\|_{(n)} = 1\}} \|Ax\|_{(m)}$.

[3] These errors will depend on the choice of norm. Higham (1996) refers to the numerator as the 'forward error' and the denominator as the 'backward error'.

[4] Golub and Van Loan (1996, 80-2) use calculus to discuss the sensitivity of x to perturbations in A and b and its relation to the condition number when A is square. See also chapter 7 of Higham (1996).

mismeasurement. For a square matrix A the condition number is $cond(A) = \|A\| \cdot \|A^{-1}\|$.

The elasticity of interest (5) is bounded by $cond(A)$:[5]

$$\frac{\|\delta x\|}{\|x\|} \div \frac{\|\delta b\|}{\|b\|} \leq cond(A) \tag{6}$$

In a linear system the identity matrix, or a scalar a transform thereof, has the smallest possible condition number from induced norms; the condition number is 1.[6] The condition number is thus a measure of 'closeness' to the identity matrix ideal.

3.2 Perturbations in A

Suppose that the data mismeasurement is in A rather than b in equation (1). As Trefethen and Bau (1997, 95) show,

$$\frac{\|\delta x\|}{\|x\|} \div \frac{\|\delta A\|}{\|A\|} \leq \left\|A^{-1}\right\| \|A\| = cond\,(A) \tag{7}$$

Thus, the effect of mismeasurement in A on the computed solution for x still depends on the condition of A. And when A is square and nonsingular, the relative condition number of the problem is simply $cond(A)$.

3.3 Perturbations in x

Perturbations in x have the same condition number as perturbations in b. This can be derived by noting that $x = A^{-1}b$, which is mathematically equivalent to $b = Ax$, except with A^{-1} replacing A (and with the vectors switched). But we know that the condition number of this problem is as before: $\|A\| \|A^{-1}\|$.

[5] See Trefethen and Bau (1997).

[6] An orthogonal basis, such as $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ will also have a condition number of 1, but such a basis could not arise in typical econometric applications.

3.4 Choosing norms to obtain condition numbers

Ideally we would like to obtain bounds that are independent of the choice of norm, but such results are not generally available. For the L_2 norm $cond\,(A) = \frac{\sigma_1}{\sigma_n}$ where σ_1 is the largest singular value of A and σ_n is the smallest singular value from the singular value decomposition of A (Trefethen and Bau 1997, 95). In practical applications Judd (1999, 68) proposes using the ratio of largest to smallest eigenvalues – the spectral condition number – in place of $cond\,(A)$. However, Golub and van Loan (1996) note that the largest and smallest eigenvalues are bounded by the largest and smallest singular values, which suggest that the ratio of singular values may be a better choice for calculating a condition number for a problem.[7]

Return momentarily to the computational problem associated with digital computing. Suppose that $cond(A) = k \times 10^j$. Then as a numerical rule of thumb, j digits of precision are lost in calculating x. Since we only want a rough idea of the magnitude of j the exact choice of norm used to calculate the condition number may not matter. Using Hilbert matrices, which are known to be ill-conditioned as the dimension of the matrices increases, Judd (1999, 69) illustrates that there is a large degree of agreement on j for the condition number from the matrix norm induced from $\|x\|_\infty = max_{1\geq i\geq m}|x_i|$ and the spectral condition number. Similarly, for the Hilbert matrices the value of j from the ratio of eigenvalues and the ratio of singular values only differ by at most one digit, and that occurs when j is around 17.[8]

4 Hessians and condition numbers

Consider Hessians in the context of the linear model equation (2) above. The Hessian is the matrix of the second derivatives of the objective function (say the sum of squared errors) relative to the parameters. Ie let the sum of squared errors, SSE be the following function:

$$\Sigma_{i=1}^{T}(\epsilon_i)^2 = SSE = f\,(y, W; \gamma)$$

[7] The Gauss `cond()` command is the ratio of largest and smallest singular values, not the ratio of eigenvalues.

[8] Using double precision arithmetic in `Gauss` to calculate these ratios.

Then the first order condition (the gradient) is:

$$\frac{\partial f(\cdot)}{\partial \gamma} = FOC = W'y - W'W\gamma$$

and the Hessian is:

$$\frac{\partial^2 f(\cdot)}{\partial \gamma^2} = -W'W = -A$$

But we know that, for numerical reasons, we want A to be as close to an identity matrix as possible. Thus, we want the *Hessian* to be close to an identity matrix (the minus sign can be thought of as a redundant scalar). That is, we want the Hessian to have a small condition number that is close to 1. Since the Hessian in the linear regression model is $W'W$, we (ideally) want the columns in W to be orthogonal and we want the diagonal elements of $W'W$ to have the same scale. Assuming that the columns of W have the same means, then we also require that the columns of W have (approximately) the same variances. This requirement parallels the desire for independent variation in experimental design (so that individual effects can be established).

In a maximum likelihood context, the inverse of the expected value of the negative of the Hessian provides the Cramer-Rao lower bound on the variance-covariance matrix of the parameter estimates. There is thus a natural connection between the variance-covariance matrix of the parameters, the Hessian, and the conditioning of the maximum likelihood problem. Large variances imply imprecise parameter estimates. The arrival of new data, or data mismeasurement, can lead to substantial change in parameter estimates if the variance-covariance matrix is ill-conditioned.

5 Hessians in nonlinear optimisation: An example

The Gauss-Newton method for finding the minimum of a nonlinear function provides another example that illustrates the importance of the Hessian (see Davidson and MacKinnon 1993, 201-3 for discussion of this maximisation algorithm). Suppose that we have a nonlinear function $Q(\theta)$ and we want to minimise or maximise the function. One approach is to pick a starting

parameter vector $\theta^{(1)}$ and take a second order Taylor series approximation of $Q(\theta^*)$ around $\theta^{(1)}$, where θ^* is the argmax of $Q(\theta)$. Ie,

$$\begin{aligned} Q(\theta^*) &\simeq Q\left(\theta^{(1)}\right) + g\left(\theta^{(1)}\right)' \cdot \left(\theta^* - \theta^{(1)}\right) \\ &\quad + \frac{1}{2}\left(\theta^* - \theta^{(1)}\right)' H\left(\theta^{(1)}\right)\left(\theta^* - \theta^{(1)}\right) \end{aligned} \tag{8}$$

The first order condition of the RHS with respect to θ^* is:

$$g\left(\theta^{(1)}\right) + H\left(\theta^{(1)}\right)\left(\theta^* - \theta^{(1)}\right) = 0$$

But, self-evidently, this is just a linear function as in equation (1), with

$$\begin{aligned} b &= -g\left(\theta^{(1)}\right) + H\left(\theta^{(1)}\right)\theta^{(1)} \\ A &= H\left(\theta^{(1)}\right) \\ x &= \theta^* \end{aligned}$$

Since the RHS of (8) is only an approximation to the LHS, except where the function is indeed quadratic, there are implicitly higher order terms that are functions of θ^*, say $c(\theta^*)$. Thus, the term b is misspecified, since it should include $c'(\theta^*)$. Therefore the function evaluated at θ^* is unlikely to minimise $Q(\theta)$ exactly. Consequently, we set $\theta^{(2)} = \theta^*$ and continuing iterating (replacing $\theta^{(1)}$ with $\theta^{(2)}$, $\theta^{(3)}$, etc). To minimise the susceptibility of our parameter estimates at each iteration to the mismeasurement $c'(\theta^*)$, we want the Hessian evaluated at each $\theta^{(i)}$ for $i = 1, 2, \ldots$ to have condition numbers that are close to 1.

6 Summary

This note illustrates why the condition number of the Hessian is so crucial when optimising functions. In analytical optimisation techniques the condition number of the Hessian indicates the susceptibility of a parameter estimate to data mismeasurement or indeed to the arrival of new information. With digital computers numbers are approximated with a finite number of digits and hence all numerical problems are susceptible to data mismeasurement. In numerical optimisation a 'poorly conditioned Hessian' implies

that one will not get very accurate estimates of $\theta^{(2)}, \theta^{(3)}, ...$ and this means that numerical optimisation algorithms may take a long time to converge to solutions, if they converge at all. Given the importance of conditioning in numerical maximisation methods, it is likely to be desirable, when considering starting values, to calculate the condition number of the relevant Hessian to determine whether the start value is in fact reasonable. The Hessian is of course a function of both the parameters and the data. In some cases condition numbers can be improved by ensuring that both the mean and variances of the underlying data dimensions are of similar scale. If the Hessian is found to be ill-conditioned, then one needs to consider whether it is possible to rescale the data to improve the properties of the Hessian. To understand how such transformations might be implemented, analytical representations of the Hessian are likely to be highly desirable.

References

Davidson, R and J G MacKinnon (1993), *Estimation and inference in econometrics*, Oxford University Press, New York.

Golub, G H and C F V Loan (1996), *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 3rd edn.

Higham, N J (1996), *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Judd, K L (1999), *Numerical Methods in Economics*, MIT Press, Cambridge, Mass.

Trefethen, L N and D Bau, III (1997), *Numerical Linear Algebra*, Society for Industrial and Applied Mathematics, Philadelphia.