KM 26(1987)
pag 5 - 12

THE NUMERICAL ERROR IN NONLINEAR PARAMETER ESTIMATION

Arie ten Cate *)

Abstract

The iterative nonlinear estimation of the parameters of a statistical model is studied. The numerical error of the parameter estimates is considered from a statistical point of view: this numerical error is linked to the standard error of the parameters, via the convergence criterion of the objective function.

1. Introduction

The estimation of the parameters of a nonlinear statistical model requires a numerical optimisation: the value of some objective function of the parameters is maximised (or minimised) in a series of iterations. In this paper we will consider two types of objective functions: the log likelihood (to be maximised) and the residual sum of squares (to be minimised).

Since the number of iterations is necessarily finite, there will usually be a (small) difference between on the one hand the optimal parameters values and the function value, and on the other the computed values of the parameters and the function. Let us call the modulus (absolute value) of such a difference the *numerical error*; both the parameter values and the function value have a numerical error.

---

*) Central Planning Bureau
Department of Applied Mathematics and Computer Science
Van Stolkweg 14
2585 JR The Hague
tel. 070-514151

For the sake of simplicity, the finite accuracy of the computer, which may be another source of numerical error, is ignored here.

In this note, the idea is considered to link the numerical error of the parameters to their statistical accuracy, in particular their standard error. Consider for example a parameter which can be estimated with a very high statistical accuracy; let us say with a relative standard error of 10^{-4}. It makes sense to compute the estimate of such a parameter with more than four decimal digits precision. On the other hand, if the standard error of a parameter is of the same order of magnitude as the parameter itself (which is quite common in the social sciences), then such a precision is nonsensical.

In the sequel it is shown how this idea can be made operational. Results for the two types of objective functions mentioned above - the log likelihood and the residual sum of squares - are derived in sections 4 and 5 respectively. Each of these sections has a verbal conclusion. The results are based on a lemma which is presented in section 3, after a discussion of convergence and convergence criteria in section 2. Section 6 gives a discussion of the results. In an appendix a proof of the lemma of section 3 is given.

Interested readers who are not familiar with nonlinear estimation are referred to Goldfeld and Quandt (1972), Bard (1974), and chapter 6 of Judge *et al* (1985).

2. Convergence

With most optimising computer programs, the user can control the number of iterations, and hence influence the numerical error of the result, through *convergence criteria*. A convergence criterion may be applied to the objective function, or the parameters, or both. A convergence criterion for the parameters usually applies to all parameters uniformly.

For instance, the user may require that the iterative process continues until the first four decimal digits of the parameters are no longer changing from one iteration to the next. This is implemented in the software by testing at the end of each iteration if the relative changes of all parameters are less than some (small) user defined number - in this case 10^{-4}. One hopes then that the parameters are correct in about four decimal digits. Or, in other words, one hopes that their relative numerical error is about 10^{-4}.

There is little theory about the relation between convergence criteria and numerical errors. Further discussion of this problem is, however, beyond the scope of this paper: whether or not the results presented below are used, this problem exists. Here, we will assume that, through the convergence criterion of the objective function, the user can control the numerical error of the objective function. As we shall see, this leads in turn to control of the numerical errors of the parameters.

3. A lemma

Throughout this note, it will be assumed that the objective function around its top (or bottom) can be adequately approximated by a quadratic function. The following lemma is about the behaviour of such a function. It shows how a deviation of a quadratic objective function from its maximum (or minimum) value can be translated to a deviation of the parameters from their optimum value, as follows.

Let $F(\theta)$ be a quadratic function of a real vector θ; say

$$F(\theta) = a - (\theta-b)'H(\theta-b)/2 \ . \tag{1}$$

Here a, b, and H are a scaler, a vector and minus the Hessian matrix respectively. Let H be symmetric and positive definite, from which it follows immediately that $F(\theta)$ reaches the maximum $F(\theta)=a$ for $\theta=b$. Then, for every real positive c the following holds. If

$$a - F(\theta) \leq c \tag{2}$$

then for every element θ_i of θ:

$$|\theta_i - b_i| \leq \sqrt{2c(H^{-1})_{ii}} \ . \tag{3}$$

A proof of (3) is given in the Appendix.

4. Maximum likelihood

In this section the maximum likelihood estimation of the parameter vector θ of

a nonlinear model is discussed. It is assumed that the log likelihood function around the optimum, given the observed data, can be adequately approximated by a quadratic function. In most cases, this gives a better approximation than approximating the likelihood function itself (a non-negative function) with a quadratic. Also, think of the standard linear regression model with normally distributed disturbances, where the log likelihood function is everywhere exactly quadratic.

Let the log likelihood function near its maximum be given by the right hand side of (1). Then $\theta=b$ gives the maximum likelihood estimate of θ. The idea discussed in the introduction can now be written as

$$|\theta_i - b_i| \leq s_i f \ , \tag{4}$$

for all elements θ_i of the vector θ. Here f is some small positive number which relates the numerical error of the parameter θ_i to its estimated standard error s_i.

The s_i values can be derived from the well known theorem that under certain regularity conditions the inverse of H in the optimum is a consistent estimator of the covariance matrix of the parameter estimates: $s_i^2=(H^{-1})_{ii}$. Substitution into (4) gives

$$|\theta_i - b_i| \leq f\sqrt{(H^{-1})_{ii}} \ . \tag{5}$$

The lemma of section 3 shows that condition (5) is met if the distance between the log likelihood and its maximum is not greater than

$$c = f^2/2 \ . \tag{6}$$

In words: the numerical error of the parameters in a nonlinear maximum likelihood problem will be less than or equal to a fixed fraction of their estimated standard error - this fraction being the same for all parameters - if we put the numerical error of the log likelihood function less than or equal to one half of the square of this fraction.

Notice finally that the *absolute* numerical error of the log likelihood function is at stake here (and the relative numerical error of the likelihood function itself). Since usually the log likelihood is maximised (not the likelihood itself), the application of the result given above requires an absolute func-

tion convergence criterion in optimising software. As far as I know the available standard optimising software (in the NAG and IMSL libraries), only a relative criterion is allowed for.

5. Least squares

Consider the following nonlinear regression model, in conventional notation:

$$y_t = f(\theta; x_t) + u_t \quad , \tag{7}$$

for t=1,..,n. The disturbances u_t are stochastic variables. They are independently and identically distributed, with expectation zero. Conditions have been given in the literature under which the least squares estimator of θ is consistent, both with and without the specification of the form of the distribution of the disturbances u_t. In the first case, if the disturbances are normally distributed, then the least squares estimator is the maximum likelihood estimator; in the second case there is no likelihood function. In both cases, the covariance matrix of the least squares estimator of θ is consistently estimated by

$$2s^2H^{-1} \tag{8}$$

where s^2 is a consistent estimator of the variance of u_t (for all t), and H is here the Hessian matrix of the residual sum of squares in the optimum.

We now proceed along the same lines as in the previous section. It is assumed that the residual sum of squares near the optimum can adequately be approximated by a quadratic function. Let minus the right hand side of (1) be this function.

Next, consider again the inequality (4). We replace s_i in (4) by the consistent estimate according to (8):

$$|\theta_i - b_i| \leq sf\sqrt{2(H^{-1})_{ii}} \quad . \tag{9}$$

The lemma of section 3 shows that condition (9) is met if the distance between the sum of squares and its minimum is not greater than

$$c = s^2f^2 \quad . \tag{10}$$

The usual s^2 estimator is the residual sum of squares divided by n-k, where k is the number of parameters. Then we have a relative numerical error of the sum of squares equal to

$$f^2 / (n-k) . \qquad (11)$$

In words: the numerical error of the parameters in a nonlinear least squares problem will be less than or equal to a fixed fraction of their standard error - this fraction being the same for all parameters - if we put the relative numerical error of the least squares function less than or equal to the square of this fraction divided by n-k, the denominator of the estimator of the error variance.

6. Discussion

In this paper, the numerical errors (the errors due to the numerical optimisation procedure) in the estimation of nonlinear statistical models have been studied from a statistical point of view. The results are very simple and applicable to all maximum likelihood problems (if the software has an absolute convergence criterion, in stead of a relative one) and to most least squares problems.

As far as I know, the results in sections 4 and 5 are new. However, it must be noted here that Cramer (1986, p. 73) also suggests to link the numerical error in nonlinear maximum likelihood problems to the standard error of the parameters - though he presents no way to do so.

The method presented here is most useful in cases where the cost of an iteration is very high (for instance with numerical procedures within an iteration, or with a vary large amount of data), and the relative standard error differs greatly between the parameters. In such cases, the usual approach - with one convergence criterion for all parameters - may lead to a waste of computer time: for the parameters with a large standard error, the criterion is too sharp. A convergence criterion with respect to the objective function, as presented in the previous two sections of this paper, does not have this defect.

Appendix. Proof of the lemma of section 3

This proof of the lemma in section 3 is due to F.J.H. Don. (My original proof was about three times as long.)

Since the matrix H in (1) is symmetric and positive definite, inequality (2) describes a solid elipsoid. The extreme points of the ellipsoid in the direction of, say, θ_i can be derived from the first order conditions for a maximum or a minimum of θ_i under the restriction of (2), written as an equality:

$$(\theta-b)'H(\theta-b)/2 - c = 0 . \tag{A1}$$

The relevant Lagrangian is

$$L = \theta_i + m[(\theta-b)'H(\theta-b)/2 - c] = e_i'\theta + m[(\theta-b)'H(\theta-b)/2 - c] . \tag{A2}$$

Here, m is the Lagrange multiplier and e_i is the i^{th} unit vector. Then

$$dL = e_i'(d\theta) + m(d\theta)'H(\theta-b)/2 + m(\theta-b)'H(d\theta)/2 + (dm)[(\theta-b)'H(\theta-b)/2 - c] . \tag{A3}$$

Requiring $dL=0$ for any $d\theta$ and dm gives the desired first order conditions: equation (A1) and

$$e_i + mH(\theta-b) = 0 . \tag{A4}$$

Equation (A4) implies

$$\theta-b = -(1/m)H^{-1}e_i . \tag{A5}$$

Substitution of (A5) into (A1) gives

$$(1/m^2)e_i'H^{-1}HH^{-1}e_i/2 - c = (1/m^2)e_i'H^{-1}e_i/2 - c = 0 \tag{A6}$$

or

$$1/m^2 = 2c/(e_i'H^{-1}e_i) . \tag{A7}$$

Then, with (A5) and (A7), we have

$$(\theta_i - b_i)^2 = [e_i'(\theta - b)]^2 = (1/m^2)(e_i'H^{-1}e_i)^2$$
$$= 2c(e_i'H^{-1}e_i) = 2c(H^{-1})_{ii} . \qquad \text{(A8)}$$

Taking the square root of the first and last member of (A8) gives

$$|\theta_i - b_i| = \sqrt{2c(H^{-1})_{ii}} . \quad /// \qquad \text{(A9)}$$

References

Bard, Y., 1974, Nonlinear parameter estimation. Academic Press, New York.

Cramer, J.S., 1986, Econometric applications of maximum likelihood problems. Cambridge University Press, Cambridge.

Goldfeld, S.M. and R.E. Quandt, 1972, Nonlinear methods in econometrics. North-Holland, Amsterdam.

Judge, G.G., W.E. Griffiths, R. Carter Hill, H. Luetkepohl, and T.-C. Lee, 1985, The theory and practice of econometrics (second edition). Wiley, New York.

Ontvangen: 13-02-1987
Geaccepteerd: 29-06-1987