Control
and Cybernetics
VOL. 10 (1981) No. 1–2

On the alternative forms of functionals and their derivatives for a class of parameter estimation problems

by

ROMAN WEINFELD
Polish Academy of Sciences
Systems Research Institute
Warszawa, Poland

In the paper we consider the problem of least squares estimation of nonlinear model in which the parameters to be estimated can be reclassified as linear-nonlinear.

Variable projection method is presented and another version of it is proposed.

Short numerical discussion is given.

1. Introduction

In fitting nonlinear models by least squares the estimates of the parameters cannot be in general determined analytically. Therefore the usual way to obtain them is to employ one of the several existing iterative algorithms. Thus by making an initial guess to the optimum parameter values and then by consecutive repetitions of the iterative step, given an appropriate stopping rule, we can obtain estimates that have approximately the least squares property. However, when the number of unknown parameters is large the required computer time may appear unduly great. This is the reason why many authors have focused their attention on methods exploiting the properties of some particular classes of models. More specifically the majority of papers is devoted to nonlinear models in which some of the parameters are linear, [2–5] and others.

In the present paper special attention would be paid to the method first introduced by Scolnik [7] and then extended to a wider class of models by Guttman, Pereyra, Scolnik [2]. This double-step algorithm proceeds with determining first the least squares estimates of nonlinear parameters and then it makes use of linear regression to produce the estimates of linear parameters.

The aim of this paper is to show that the variable's projection method can be extended to a wider class of models and that to compute functional which is to be minimized (and its derivatives) stable numerical algorithms can be employed.

In Section 2 a general problem of nonlinear estimation is formulated. In Section 3 the variable projection method is presented and then a theorem given in [2] on the equality between estimates is reformulated. Finally Section 4 gives a formula for derivatives which is suitable for numerical handling.

2. The problem of nonlinear regression

Let us consider a system of L inputs $\boldsymbol{x}^T=[x_1, ..., x_L]$ and one output y, subjected to stochastic disturbance z

$$y=\tilde{f}(\boldsymbol{x}, z) \tag{1}$$

Let us assume that a set of N independent observations of input and output values is given

$$\{(\boldsymbol{x}_n, y_n),\ \boldsymbol{x}_n^T=[x_{1,n}, ..., x_{L,n}],\ n=1, ..., N\} \tag{2}$$

and moreover that for each observation

$$\begin{aligned} \hat{y}_n&=f(\boldsymbol{x}_n, \boldsymbol{p}) \\ y_n&=\hat{y}_n+w_n \end{aligned} \qquad n=1, ..., N \tag{3}$$

where f in (3) is a mathematical model of (1), depending on $2K$ unknown parameters $\boldsymbol{p}^T=[p_1, ..., p_{2K}]$ and at least some of them enter into f nonlinearly. $\boldsymbol{x}_n, y_n, w_n$ are respectively input, output and disturbance for n-th observation.

Since disturbances are usually unknown it is in general impossible to determine the exact form of f. Our aim is therefore to find parameters $\boldsymbol{p}^*$ of f which minimize the following sum of squares

$$s(\boldsymbol{p})=\|y-\hat{y}\|^2=\sum_{n=1}^{N} w_n^2=\sum_{n=1}^{N}(y_n-f(\boldsymbol{x}_n, \boldsymbol{p}))^2 \tag{4}$$

$$s(\boldsymbol{p}^*)=\min_{\boldsymbol{p}} s(\boldsymbol{p})$$

In the paper we consider a more particular case of estimation parameters $\boldsymbol{a}$, $\boldsymbol{b}$ appearing in models of the form

$$f(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b})=\sum_{i=1}^{K} a_i f_i(\boldsymbol{x}, \boldsymbol{b}) \tag{5}$$

where each f_i is a given nonlinear function of $\boldsymbol{b}$ $(i=1, ..., K)$.

The aim is, as previously, to determine parameters $a^T=[a_1, ..., a_K]$ and $b^T=[b_1, ..., b_K]$ which minimize (4) on the basis of observations (2), that is

$$s(a, b)=\sum_{n=1}^{N}(y_n-f(x_n, a, b))^2 \tag{6}$$

$$s(a^*, b^*)=\min_{a' b} s(a, b)$$

This problem for a more particular model

$$f(x, a, b)=\sum_{i=1}^{K} a_i f_i(x, b_i) \tag{7}$$

was investigated in [2]. We shall now resummarize the method developed there.

3. Double step variable projection algorithm

Referring to the model (7) notice that for fixed parameter values, the functions f and f_i $(i=1, ..., K)$ taken at the points $x_1, ..., x_N$ may be identified with N—dimensional vectors, i.e.

$$f^T=[f(x_1, a, b), ..., f(x_N, a, b)]$$

$$f_i^T=[f_i(x_1, b_i), ..., f_i(x_N, b_i)], \quad i=1, ..., K$$

Assume that for fixed b vectors f_i are linearly independent. This implies that $N \geqslant K$. For fixed h and changing a the vector f would then form (according to (7)) a hyperplane spanned by $(f_1, ..., f_K)$ which we will denote from now on by $\Omega(b)$. Let $P(b)$ be an orthogonal projector onto $\Omega(b)$. Consider the functional

$$r_1(b)=\|(I-P(b))\, y\|^2 \tag{8}$$

Suppose, we minimize (8) to find estimates $\hat{b}$

$$r_1(\hat{b})=\min_b r_1(b)$$

Estimates $\hat{a}$ can now be found by substituting $\hat{b}$ into (7) and applying linear regression i.e. by minimizing

$$r_2(a)=\|y-\sum_{i=1}^{K} a_i f_i(\hat{b}_i)\|^2 \tag{9}$$

It turns out that estimates b^* are equal to the estimates $\hat{b}$:

THEOREM 1. [2]. *If the functionals* (6), (8) *have unique minima then*

$$(a^*, b^*)=(\hat{a}, \hat{b})$$

To complete the method let us now recall (see [2]) the formula for derivative of a projection functional. This method enables us to apply gradient minimization algorithms

$$\frac{\partial P_y}{\partial b_i}=\frac{1}{\|\hat{e}_{(i)}\|}\left[(\psi_i^T y)\, e_{(i)}+(y^T e_{(i)})\,\psi_i\right] \tag{10}$$

where

$$\psi_i=\frac{\partial f_i(b_i)}{\partial b_i}-P\frac{\partial f_i(b_i)}{\partial b_i}$$

$$\hat{e}_{(i)}=f_i-P_i f_i, \qquad e_{(i)}=\frac{\hat{e}_{(i)}}{\|\hat{e}_{(i)}\|}$$

P_i — projection on $(f_1, \dots, f_{i-1}, f_{i+1}, \dots, f_K)$, $i=1, \dots, K$.

Finally notice that although this method applies to the model (7), it can be easily extended to the model (5). However this would need a bit more general formula for derivatives.

4. Calculations of derivatives

The algorithm just described has evident advantages. However the number of multiplications necessary for calculating the value of functional (8) and its derivatives as given above is large $(O(NK^3))$. For large N and K it annihilates to great extent the advantages of reducing the number of parameters. Hereafter we would show how the derivative of (8) can be easily computed, what can considerably simplify an algorithm.

First we shall prove the following theorem:

THEOREM 2. *Let W be $N\times K$ matrix $(K<N)$, the columns of which are formed by K linearly independent N—dimensional vectors $f_1, \dots, f_K$. Let W^* denote a $N\times(K+1)$ matrix, the columns of which are formed by vectors $f_1, \dots, f_K, y$. If $U=W^T W$ and $U^*=W^{*T} W^*$ then*

$$r_1(b)=\frac{\det U^*}{\det U} \tag{11}$$

Proof. If $\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}$ is a positive definite matrix then obviously

$$\det\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}=\det\begin{pmatrix} A & B \\ O & C-B^T A^{-1} B \end{pmatrix}=\det A \det(C-B^T A^{-1} B) \tag{12}$$

In our case

$$A=W^T W, \quad B=W^T y, \quad C=y^T y$$

Hence

$$\frac{\det U^*}{\det U}=y^T y-y^T W(W^T W)^{-1} W^T y$$

It is known that $P=W(W^T W)^{-1} W^T$ is as ymmetric projection on $(f_1, ..., f_K)$ that is

$$P^2=P \text{ and } P^T=P$$

so we have

$$\|(I-P)y\|^2=y^T(I-P)^T(I-P)y=y^T y-y^T Py=y^T y-y^T W(W^T W)^{-1} W^T y$$

this completes the proof.

Since our aim is to minimize $r_1(b)$ we can assume that $\det U^* \neq 0$.

Before we shall state the next theorem let us assume that:

a) if A is $M_1 \times L_1$—matrix, B is $M_2 \times L_2$—matrix then both multiplication or addition of A and B can be done after appropriate extension of A and/or B with zero rows or columns;

b) derivative of matrix A is a matrix containing the derivatives of the elements of A.

Now we shall show that

$$\frac{\partial r_1}{\partial b_i}=2z^T\left(\frac{\partial W^*}{\partial b_i}\right)^T W^* z \tag{13}$$

where

$$z=(W^T W)^{-1} W^T y-e_{K+1}$$

$$e^T_{K+1}=[0, ..., 0, 1] \quad e_{K+1} \in R^{K+1}$$

but first let us recall some well known facts.

Fact 1 [1]

$M-P$ pseudoinverse of the matrix A which we shall denote by A^+ has the following properties:

a) if A $N \times K$—matrix $(N>K)$ has rang K then $A^+=(A^T A)^{-1} A^T$;

b) $(A^T A)^+=A^+(A^T)^+$;

c) $(A^T)^+=(A^+)^T$;

d) AA^+ is a projection operator on a space spanned by columns of A;

e) if x is a vector then $x^+=\frac{x^T}{\|x\|^2}$.

Fact 2 [1]

Let A $N \times K$—matrix, $c \in R^N$, $d \in R^K$ if $g=A^+ c$, $h=d^T A^+$, $u=(I-AA^+)c$

$$v=d^T(I-A^+A), \ \beta=1+d^T A^+ c$$

then if $u \neq 0$, $v \neq 0$

$$(A+cd^T)^+=A^+-gu^+-v^+h+\beta v^+h^+$$

Fact 3

If A $K\times K$—matrix has rank K then

$$\frac{\partial \det A}{\partial b}=\det A\cdot tr\left(A^{-1}\frac{\partial A}{\partial b}\right)$$

Let Df stand for $\partial f/\partial b$. We can formulate the following theorem:

THEOREM 3. *Let W, W^*, U, U^* be defined as in Theorem 2. If $r_1=\|(I-WW^+)\,y\|^2$* then

$$Dr_1=2z^T(DW^*)^T\,W^*\,z$$

where

$$z=W^+\,y-e_{K+1},\ e_{K+1}^T=[0,\dots,0,1],\ e_{K+1}\in R^{K+1}$$

Proof. By Theorem 2 we have $r_1=\dfrac{\det U^*}{\det U}$. Differentiating this ratio we have (by Fact 3)

$$Dr_1=r_1\,tr\,(U^{*-1}\,DU^*)-r_1\,tr\,(U^{-1}\,DU)=r_1\,tr\,(U^{*-1}\,DU^*-U^{-1}\,DU)\qquad(14)$$

We would like to show that $U^{*-1}=U^{-1}+C$. This together with observation that

$$tr\,(U^{-1}\,DU^*-U^{-1}\,DU)=0$$

would reduce (14) to the following

$$Dr_1=r_1\,tr\,(C\,DU^*)\qquad(14a)$$

For that purpose let us notice that

$$W^*=W+ye_{K+1}^T$$

and one can easily verify the assumptions of Fact 2 hence

$$W^{*+}=(W+ye_{K+1}^T)^+=W^+-\boldsymbol{g}\boldsymbol{u}^+-\boldsymbol{v}^+\,\boldsymbol{h}+\beta\boldsymbol{v}^+\,\boldsymbol{u}^+\qquad(15)$$

but

$$\boldsymbol{h}=e_{K+1}^T\,W^+=0\ \text{ so }\ \beta=1\ \text{ and }\ \boldsymbol{v}=e_{K+1}^T$$

and (15) can be rewritten

$$W^{*+}=(W+ye_{K+1}^T)^+=W^+-(g-e_{K+1})\,\boldsymbol{u}^+$$

Notice now, that $0=W^+\,\boldsymbol{u}=\boldsymbol{u}^+W^{*T}$ and $\|\boldsymbol{u}\|^2=r_1$ thus by Fact 1

$$U^{*-1}=U^{-1}+\frac{1}{r_1}(g-e_{K-1})\,(g-e_{K+1})^T.\qquad(16)$$

If we call $z=g-e_{K+1}=W^+\,y-e_{K+1}$ then from (14a) and (16) we have

$$Dr_1=tr\left((g-e_{K+1})\,(g-e_{K+1})\right)^T DU^*=z^T\,DU^*\,z=$$
$$=z^T\,(DW^{*T}\,W^*+W^{*T}\,DW^*)\,z=2z^T\,DW^{*T}\,W^*\,z\qquad\blacksquare$$

5. Remarks

In the present section we would like to stress the necessity of correct implementation of the presented formulas. Depending on the problem one should choose an algorithm which is as fast and exact as possible and which has minimal storage requirements. Below we would give some remarks concerning the existing algorithms which can be employed during evaluating r_1 and its derivatives.

Functional r_1 can be evaluated by solving normal equations or directly from (11). However, it is usually recommended to make use of orthogonalization methods such as Householder method [8] or modified Gram-Schmidt method [9]. Errors resulting from these methods are proportional to the condition number of the matrix W and only the coefficient in the error estimation formula of the latter is greater. Both require essentially NK^2 multiplications and $NK^2+K^2/2$ storage locations. Direct solving of normal equations gives an error which is proportional to the square of condition number of W, but the number of multiplications decreases to about $\frac{1}{2}NK^2+\frac{1}{6}K^3$ (if Cholesky's method is used [8]) and relatively small number of storage locations is required (about K^2). It is worthwile to notice that in evaluating r_1 by means of Cholesky's method there is no need to evaluate separately numerator and denominator since the value of denominator can be extracted on the last but one step of this method. Error resulting this time is also proportional to the square of condition number of W although with a little smaller coefficient that the one obtained when solving normal equations. Analoguesly, the number of multiplications which is roughly the same, would slightly differ because of different coefficients.

The computation of the derivatives requires a relatively small number of multiplications. More precisely, vector z in (13) has already been computed while determining r_1. Consequently multiplying first $W^* z$ and them $z^T(\partial W^*/\partial b_i)^T$ for $i=1, ..., K$, we can get all the derivatives in a process involving about NK^2 multiplications. In fact, in the case of model (7) the indispensable number of multiplications decreases to about NK.

We would like to mention also that $W^* z=-\mathbf{r}$ where residual vector $\mathbf{r}=(I-P)_y$ and $z^T DW^{*T}=(yW^+)^T DW^T$ hence formula (13) can be replaced by another one $Dr_1=-2rDWW^+\mathbf{y}$ which may be easier to implement.

Now we can sum up all the general remarks:

a) when the gradient optimization methods are to be used:

- if A is known to be well conditioned it is reasonable to solve the normal equations,
- otherwise it is bettwe to make use of Householder orthonalization algorithm;

b) when optimization algorithms without derivatives are to be used:

- if A is known to be well conditioned it is better to make use of formula (11).
- otherwise Householder orthogonalization algorithm should be employed.

All these directions were adopted in the computer implementation of the described method. For minimization of (8) the Davidon's procedure has been employed with necessary derivatives calculated according to (10).

The algorithm has been applied to the model

$$y=a_1\, e^{-b_1 x}+a_2\, e^{-b_2 x}$$

for two different data sets ($N=6$ and $N=72$). Both versions of the algorithm produced the same estimates while the computer time differed especially in the second test ($N=72$) and for the version presented above was 30% shorter.

References

[1] Campbell S. L., Meyer C. D. Generalized Inverses of Linear Transformations. London 1979.

[2] Guttman I., Pereyra V., Scolnik H. D. Least Squares Estimation for a Class of Nonlinear Models. *Technometrics* **15**, 2 (1973).

[3] Harville D. A. Fitting Partially Linear Models by Weighted Least Squares. *Technometrics* **15**, 3 (1973).

[4] Nelson D. L., Lewis T. O. A Method for the Solution of the Non-Linear Least Squares Problems when Some of the Parameters are Linear. *Texas J. Sci.* **21** (1970), 480.

[5] Papaioannou T., Kempthorne O. Parallel Tangents and Steepest Descent Optimization Algorithm — A Computer Implementation with Application to Partially Linear Models. *J. Statist. Comput. Simul.* **1** (1972), 349–376.

[6] Ralston A. Introduction to Numerical Analysis (Polish Translation). Warszawa 1971.

[7] Scolnik H. D. On the Solution of Nonlinear Least Squares Problems. Information Processing 71. Amsterdam 1972.

[8] Wilkinson J. H., Reinsch C. Linear Algebra. Berlin 1971.

[9] Björck A. Solving Linear Least Squares Problems by Gram-Schmidt Orthogonalization, *BIT* 7 (1967), 1–21.

Received, June 1980.

O równoważnych postaciach funkcjonałów i ich pochodnych w pewnych zadaniach estymacji parametrów

Rozpatrzono zadanie identyfikacji parametrów metodą najmniejszej sumy kwadratów w modelach częściowo liniowych o postaci

$$f(x, a, b)=\sum_{i=1}^{K} a_i\, f_i\,(x, b)$$

Opisano metodę rzutowania zmiennych Scolnika a następnie zaproponowano inną jej wersję. W wersji tej wygodna postać pochodnej umożliwia uproszczenie procesu liczenia na każdym kroku zastosowanej iteracyjnej metody minimalizacji.

Об эквивалентных видах функционалов и их производных в некоторых задачах оценки параметров

В статье рассматривается задача идентификации параметров методом наименьших квадратов в частично линейных моделях типа

$$f(x, a, b)=\sum_{i=1}^{K} a_i f_i(x, b)$$

Представлен метод проекции переменных, а затем предложен новый его вариант. В этом варианте удобный вид производной позволяет упростить процесс вычисления на каждом шаге применяемого итерационного метода минимизации.