

The initial location at the University of Vaasa, Finland  
<http://lipas.uwasa.fi/~ts/publicat/GradiHeikkilaSalmi.pdf>

Heikkilä, Matti and Timo Salmi (1974). A Gradient Method for Least Squares Estimation in Nonlinear Regression Analysis. *Finnish Journal of Business Economics* 23:4, 329-336.

This paper is reproduced at the University of Vaasa in the electronic format with the permission of The Finnish Journal of Business Economics. Copyright © 1974 by The Finnish Journal of Business Economics and the authors.

<http://lta.hse.fi/>

<http://www.uwasa.fi/laskentatoimi/english/personnel/salmitimo/>

# Liiketaloudellinen Aikakauskirja

The Finnish Journal of Business Economics

**4 - 1974**

23. VUOSIKERTA

**Osmo A. Wiio — Martti Helsilä**

Auditing Communication in Organizations: A Standard Survey  
"LTT Communication Audit"

**Klaus Kerppola**

The Profitability of Large Firms in the Finnish Forest Industry

**Matti Heikkilä — Timo Salmi**

A Gradient Method for Least Squares Estimation in Nonlinear  
Regression Analysis

**Antti Korhonen**

Term Structure of Interest Rates in Finland 1963—1973

**Veikko Leivo — Uolevi Lehtinen — Pekka Akkanen**

Huoltotarve ja etäisyys autokorjaamon valinnassa

---

Toimituskunta HUUGO RANINEN (päätoimittaja), MIKA KASKIMIES (toim.siht.), LEO AHLSTEDT,  
JAAKKO HONKO, EINO NIINI, REINO ROSSI, MARTTI SAARIO ja FEDI VAIVIO

Toimituksen Runeberginkatu 22—24, 00100 Helsinki 10. Aikakauskirja ilmestyy vuosittain neljänä  
osoite: niteenä. Tilaushinta 30:—

**The Finnish Journal of Business Economics**

Address: Runeberginkatu 22—24, 00100 Helsinki 10, Finland.

# **Liiketaloudellinen Aikakauskirja**

The Finnish Journal of Business Economics

**Special Edition 4-1974**

MATTI HEIKKILÄ — TIMO SALMI

## **A Gradient Method for Least Squares Estimation in Nonlinear Regression Analysis**

---

Toimituskunta HUUGO RANINEN (päätoimittaja), MIKA KASKIMIES (toim.siht.), LEO AHLSTEDT, JAAKKO HONKO, EINO NIINI, REINO ROSSI, MARTTI SAARIO ja FEDI VAIVIO

Toimituksen Runeberginkatu 22—24, 00100 Helsinki 10. Aikakauskirja ilmestyy vuosittain neljänä osoite: niteenä. Tilaushinta 40:—

**The Finnish Journal of Business Economics**

Address: Runeberginkatu 22—24, 00100 Helsinki 10, Finland.

# A Gradient Method for Least Squares Estimation in Nonlinear Regression Analysis

## 1. Introduction

Consider the task of finding the least squares estimates for the  $p$  parameters  $b_1, \dots, b_p$  in the specified regression model  $Y = f(X_1, \dots, X_k; b_1, \dots, b_p)$  given the  $n$  observations  $Y_i, X_{1i}, \dots, X_{ki}$  ( $i=1, \dots, n$ ). This task is accomplished by finding the values of  $b_1, \dots, b_p$  which minimize the error sum of squares

$$S(b_1, \dots, b_p) = \sum_{i=1}^n \left\{ Y_i - f(X_{1i}, \dots, X_{ki}; b_1, \dots, b_p) \right\}^2$$

One way to treat this task is to use a suitable gradient method for minimizing "the objective function"  $S(b_1, \dots, b_p)$  "subject to no constraints". However, if the regression model is linear in parameters (e.g.  $Y = b_0 + b_1X$ ) the well-known normal equations can be applied. In the nonlinear case there are several ways to treat the task, among which gradient methods form one group.

In this paper we introduce a gradient method with speeded-up convergence for the minimization.

---

For example, fitting the simple nonlinear regression model  $Y = b_1 \sin b_2 X$  given only the three observations below could be stated as the following mathematical programming problem:  $\min_{b_1, b_2} S(b_1, b_2) = (1 - b_1 \sin b_2)^2 + (3 - b_1 \sin 2b_2)^2 + (2 - b_1 \sin 3b_2)^2$ .

$$b_1, b_2$$

The observations  $(X, Y)$ : (1,1), (2,3), (3,2).  
(The fitted function is  $Y = 2.496 \sin 0.694X$ .)

## 2. General Description of the Gradient Method

Basically the gradient method to be introduced proceeds as follows. First, the "variables"  $b_1, \dots, b_p$  are given suitable initial values. Then the value of the objective function  $S(b_1, \dots, b_p)$  and its gradient  $\nabla S(b_1, \dots, b_p)$  are calculated at the initial point. The values of the partial derivatives in the gradient may be calculated from its expression, if available, or by using the well-known approximative techniques. A suitable step (steplength) is taken in the direction of the negative gradient  $-\nabla S(b_1, \dots, b_p)$  (negative, since we are minimizing the

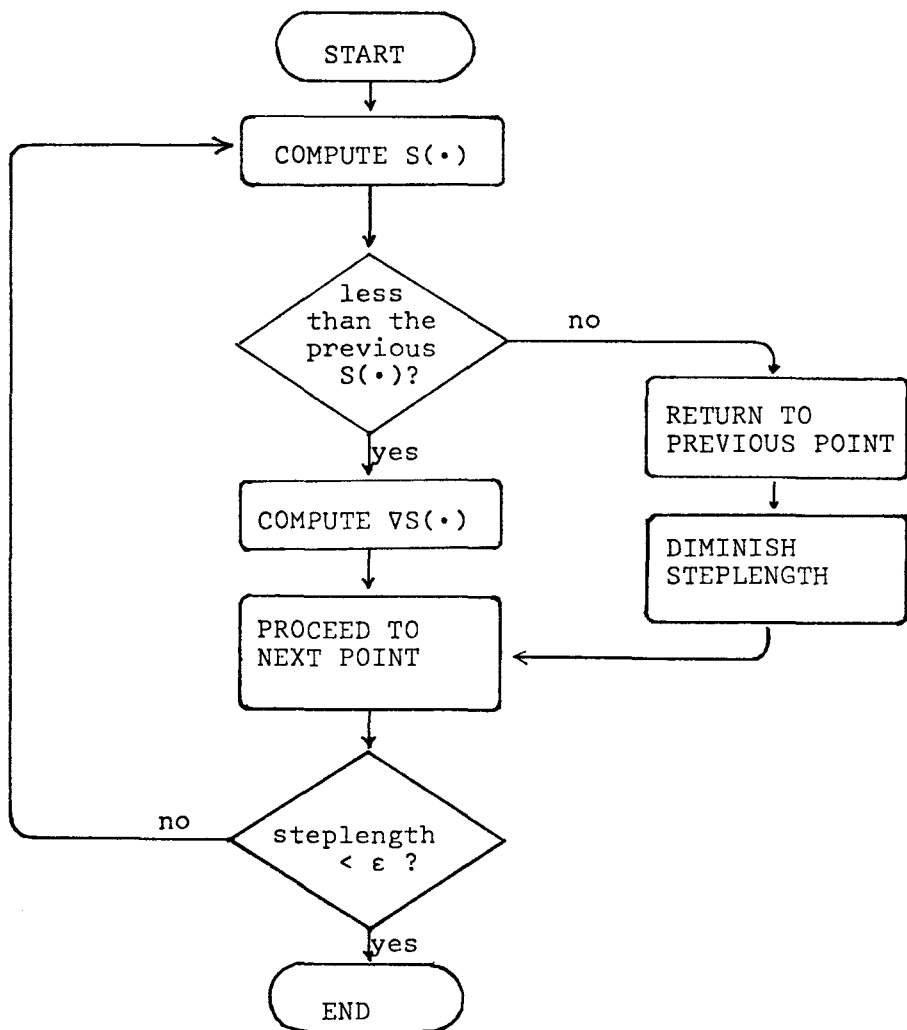


Figure 1

objective function), and thus a new point is arrived at. In the new point the value of the objective function is calculated anew. If this new value is less than the value of the objective function at the previous point the iteration procedure is continued from the new point, again by taking a step in the direction of the negative gradient, which has been recalculated at the new point. If, however, the value of the objective function is greater than before at the new point the method returns to the previous point. After that the steplength is reduced by a suitable factor and the reduced step is taken in the direction of the negative gradient. When the steplength becomes small enough iterations are terminated, and the current point is accepted as the minimum. Figure 1 gives a flow chart of the general procedure.

Figure 2 gives the contours of a function of two variables. The arrows illustrate steps of minimization towards the optimum.

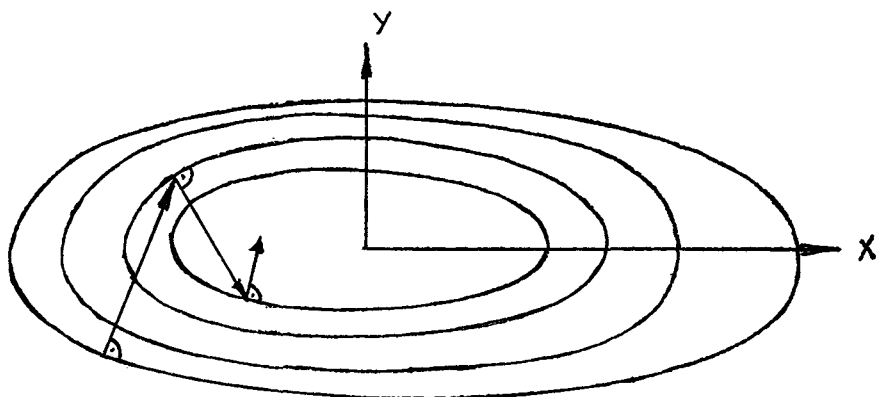


Figure 2

### 3. Exponential Smoothing to Speed up the Convergence

The gradient methods of the kind presented work properly only in simple cases. In practice several improvements have to be made. In this chapter we suggest how convergence of the method can be considerably speeded up in the case of oscillations.

Consider the following example of a function of two variables:

$$(1) \quad f(x, y) = (0.01x)^2 + (100y)^2$$

This function has its absolute minimum at the origin. Figure 3 gives a few contours of the function.

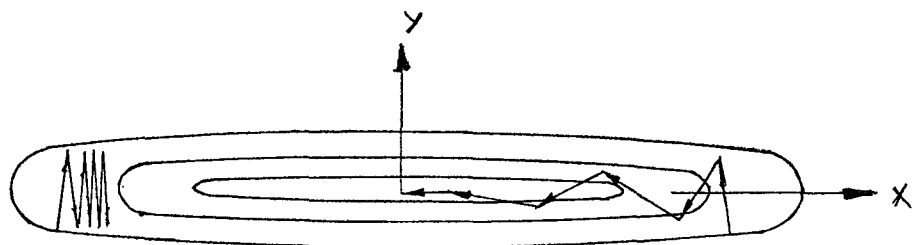


Figure 3

The arrows on the left depict the iterations with the conventional gradient method. The arrows on the right depict the iterations with speeded-up convergence, which is attained by exponentially smoothing the gradient. As is seen in Figure 3 the values of  $y$  oscillate between both sides of the "gorge", and consequently most of the steplength is wasted. This can be very easily overcome by using for each partial derivate  $\partial S(b_1, \dots, b_p) / \partial b_h$  in the gradient  $\nabla S(b_1, \dots, b_p)$  the weighted average of the present and the previous values. Denote:

(2)  $d_{hj}$  = the first partial derivate of the objective function  $S(b_1, \dots, b_p)$  of the variable  $b_h$  evaluated at the  $j$ :th iteration point.

(3)  $D_{hj}$  = the value of the exponentially smoothed first partial derivate of the objective function of the variable  $b_h$  at the  $j$ :th iteration point.  $D_{hj}$  is computed from the recursion formula (4).

(4)  $D_{hj} = \alpha d_{hj} + (1-\alpha)D_{h,j-1}$ ;  $0 < \alpha < 1$ ,  $h=1, \dots, p$ .

Next we consider some simple examples in order to demonstrate the damping effect of exponential smoothing on oscillation. If  $d_j = (-1)^j$  (the index  $h$  has been omitted for convenience) is substituted in (4) we can solve  $D_j$  from the difference equation (4) e.g. with the initial condition  $D_0 = 0$ . It is easy to see that the solution is

$$(5) \quad D_j = [\alpha / (2-\alpha)] [(-1)^j - (1-\alpha)^j].$$

When  $j$  grows  $|D_j|$  approaches the value of  $\alpha / (2-\alpha)$ . The damping depends thus on  $\alpha$ , as is seen also in Figure 4.

For a good damping effect we must pay the price of the fact that the smoothed partial derivatives follow slowly the permanent changes of the actual partial derivatives — the better the damping the slower the follow-up. For example if the sequence of the partial derivatives is 0, 1, 1, 1, ... then the solution of (4) is (7):

$$(7) \quad D_j = 1 - (1 - \alpha)^j.$$

We also see that

$$(8) \quad \lim_{j \rightarrow \infty} D_j = 1$$

and thus no error remains in the steady state. If, however, the sequence of the partial derivatives is e.g. 0, 1, 2, 3, ... then the solution of (4) is (9):

$$(9) \quad D_j = j - [(1 - \alpha)/\alpha] [1 - (1 - \alpha)^j]$$

and in the steady state an error of  $(1 - \alpha)/\alpha$  remains.

The selection of the coefficient  $\alpha$  must be made on subjective grounds. In the minimization delineated in Figure 3  $\alpha = 0.4$  was used.

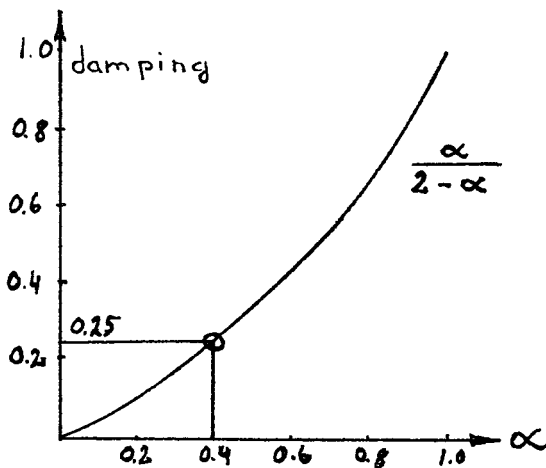


Figure 4

#### 4. Adjustment of the Steplength

Difficulties in the selection of the initial steplength and slow convergence rate after "tight bends" in the contours of the objective function are easily overcome by a simple adjustment procedure of the steplength.

Consider the following example of a function of two variables:

$$(10) \quad f(x, y) = (y - \sin 100x)^2 + (0.01x)^2.$$

This function has its absolute minimum at the origin. Figure 5 gives a few contours of the function.

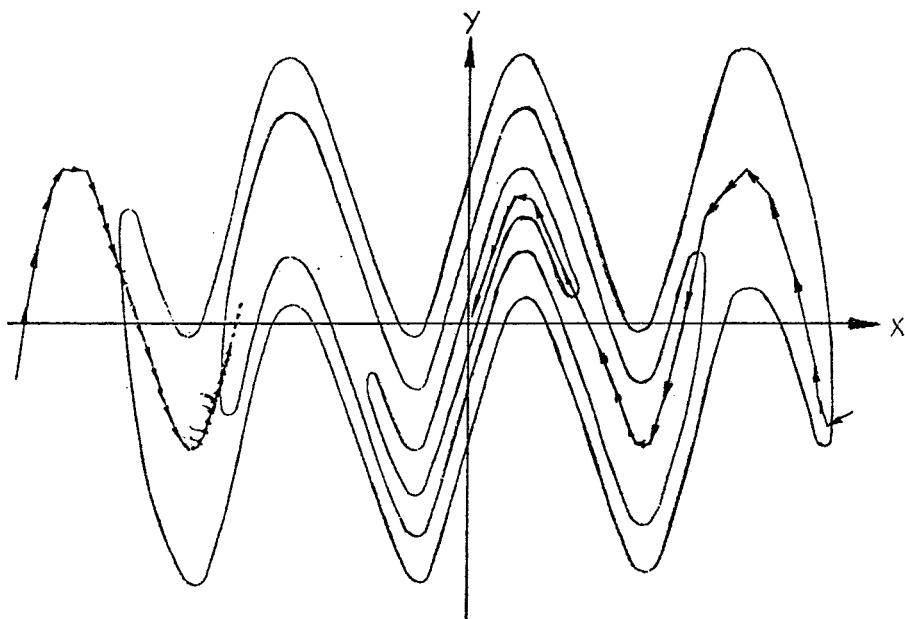


Figure 5

The arrows on the left depict the iterations with the conventional gradient method. The arrows on the right depict the iterations with a step length adjustment procedure (and exponential smoothing). On the left the step length will be reduced once or twice in each bend and if the distance to the minimum is long the computation time often becomes prohibitively long, or worse still the iterations may be terminated too early! This snag is overcome by always increasing the step length when a predetermined small number of steps have been taken without having to reduce the step length. As is seen on the right in Figure 5 the iteration procedure now gains the momentum needed to proceed to the optimum even after the bends. This adjustment procedure also quickly corrects a bad selection of the initial step length. The adjustment procedure has no obvious drawbacks and it is advisable always to include it.

## 5. Further Complications

In cases where some of the partial derivatives of the objective function  $S(b_1, \dots, b_p)$  do not decrease in absolute value in the neighbourhood of the minimum, the use of scaling-factors and an additional rule in step length adjustment can be resorted to.

Consider now the following function:

$$(11) \quad f(x,y) = (0.01x)^2 + 100/y.$$

Figure 6 gives contours of the function.

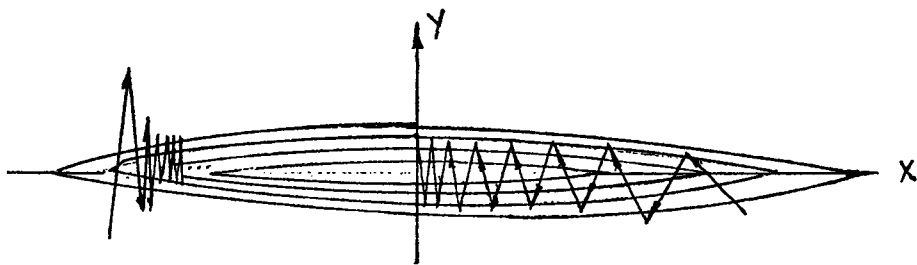


Figure 6

The arrows on the left depict iterations with exponential smoothing and step-length adjustment. Exponential smoothing does not help much in this case, since  $\partial f(x,y)/\partial y$  does not decrease on the way towards the minimum at the origin. The amplitude of the oscillations is reduced to about a fourth, but this is inadequate. This is because the oscillations in the  $y$ -direction use up the step-length almost completely. The steplength cannot be made longer, since the objective function would be increased in the value then. Fortunately, we often have a preconception of the behavior of the objective function  $S(b_1, \dots, b_p)$  in the neighbourhood of the minimum. If we know that the value of the objective function increases steeply very near the minimum for some variables, we can overcome the unnecessary oscillations by multiplying the relevant partial derivatives by special scaling-factors. For example, for function (11) we could multiply  $\partial f(x,y)/\partial y$  by the scaling-factor  $10^{-3}$ . The iterations would then proceed as shown on the right in Figure 6. Oscillation on both sides of the minimum remains but it can be damped by reducing the steplength also always when a predetermined *great* number of steps have been taken without a change in the steplength.

## 6. Choosing the Initial Value

At the beginning of an actual minimization procedure each variable  $b_1, \dots, b_p$  and the steplength must be assessed initial values.

The initial values of the variables must be chosen very carefully. This is because the success of gradient methods depends heavily on how near the minimum the initial values are chosen. In addition the objective function  $S(b_1, \dots, b_p)$  may have several local minima. In this case the initial point must lie sufficiently close to the absolute minimum, lest the iterations converge to a local minimum.

The selection of the initial steplength is not very critical. It should be made too long rather than too short in order to get started at as fast a pace as possible. It cannot be made too long, since the steplength is automatically reduced until a better point than the starting point is arrived at. If the initial steplength is too short no problems arise, although it takes a little while until the steplength adjustment procedure discussed in Chapter 4 makes the steplength sufficiently long.

## 7. Conclusion

In this paper we discussed a gradient method which is well suited for fitting nonlinear regression models by least squares minimization. We suggested a way to speed up the convergence by applying exponential smoothing for the gradient. We also discussed adjustments of steplength in order to gain momentum in the minimization and in order to damp out unnecessary oscillation, and the use of scaling-factors in special cases.

It is clear that such cases can be found where the minimization procedure discussed does not work well enough. According to our computational experience it seems, however, that in cases of nonlinear regression analysis normally arising the method discussed is highly satisfactory.