ON THE CONFLUENT APPROACH IN REGRESSION ANALYSIS

A. S. Mechenenov

We study all the methods of estimating the parameters of stochastic models of linear confluent analysis and compare the results using the examples presented.

Four figures. Bibliography: 11 titles.

1. Introduction. Multiple linear regression analysis. We begin by considering a widely known approach to the construction of estimates of unknown parameters of functional relations.

Assumption 1. A functional relation of the form

\[ y = H\delta, \; \delta \in \mathbb{R}^n, \; H \in \mathbb{R}^{n \times k}, \; \delta \in \mathbb{R}^k, \; n \geq k, \]  

(1.0)

is given, where \( H = [\eta_1, \cdots, \eta_k] \) is a known matrix, \( \delta = (\delta_1, \cdots, \delta_k)^T \) are unknown parameters, and \( \delta = (\delta_1, \cdots, \delta_n)^T \) is the unknown response vector.

Not all the quantities that occur in (1.0) are random. (We consider measured quantities to be random.) We assume that this relation holds without regard to our ability to observe it. Suppose we are able to observe (measure) only \( \delta_i, i = 1, \cdots, n \), knowing that each value of \( \delta_i \) is functionally connected by definition with the values of \( \eta_1, \cdots, \eta_k \) (neither measured nor prescribed by us, but known theoretically). Such an approach is described by the following model.

Assumption 1.1. The vector \( y = (y_1, \cdots, y_n)^T \) is a random vector of observations over a Euclidean sample space \( \mathbb{R}^n \), \( \mathcal{L}[H] \) is a linear manifold of dimension \( k \) in the space \( \mathbb{R}^n \), \( n \geq k \). We assume that there exists a set of representations of the form

\[ y = \delta_1 \eta_1 + \cdots + \delta_k \eta_k + e, \]  

(1.1)

called a linear stochastic model, where \( \eta_1, \cdots, \eta_k \) are vectors in the set \( \mathcal{L}[H] \), in which \( \eta_i \) may be a vector of \( I \)'s: \( \eta_i = (1, \cdots, 1)^T \), \( e = (e_1, \cdots, e_n)^T \) is an additive random vector of observational errors having mathematical expectation zero (\( Ee = 0 \)) and a known positive-definite covariance matrix that is independent of \( \delta \):

\[ \text{cov}(e, e) = E(e - Ee)(e - Ee)^T = Eee^T = \Sigma. \]

Problem 1.1. Knowing one observation \( \bar{y} = H\delta + e \) of a random vector \( y \), its covariance matrix \( \Sigma \), and the matrix \( H \) of full rank \( k \), estimate the true values of the parameters \( \delta \) of the model (1.1) so that the weighted quadratic form

\[ S^2 = (\bar{y} - H\delta)^T \Sigma^{-1}(\bar{y} - H\delta) \]
is minimized.

**Theorem 1.1.** The estimate \( \hat{d} \) of the parameters \( \delta \) can be computed from the system of linear algebraic equations

\[
H^T \Sigma^{-1} H \hat{d} = H^T \Sigma^{-1} y,
\]

which is called a normal equation or system of normal equations.

We have earlier discussed the method of least squares, which was proposed by Legendre [1] and Gauss [2] and which was generalized in [3] to the case of an arbitrary covariance matrix. Markov [4] and Malinvaud [5] in turn proposed entirely different approaches leading to the same results as the method of least squares.

Let us consider an example. In order better to compare the results of parameter estimation, we artificially place four points of initial data at the vertices of a square \( \{(1,1),(1,2),(2,1),(2,2)\} \), which has two axes of symmetry (the diagonal and the middle line). From these four observations we shall construct estimates of the parameters of the different functional relations. In Fig. 1.1a the model (1.1) was formed for a functional relation (1.0) of the form \( \delta = \eta \delta \), and the four measurements were approximated by the method of least squares, giving as a result the line \( \hat{y} = 0.9 \eta \), whose slope is less than 1 (regression). We process the data using the regression model with a free parameter. On the same four observations we construct an estimate for the parameters of a functional relationship of the form \( \delta = \eta \delta + \delta_1 \). In Fig. 1.2a the four measurements are approximated by the method of least squares, leading to a constant (a second axis of symmetry and even larger regression).

In regression analysis (so-named by Galton [6]) the response and only the response is subject to observational errors, since this occurs, say, in approximating a function known at \( n \) points by a set of \( k \) other