Physica A 434 (2015) 257–267

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Robust linear regression with broad distributions of errors

Eugene B. Postnikov [a,*], Igor M. Sokolov [b]

[a] Theoretical Physics Department, Kursk State University, Radishcheva st., 33, Kursk 305000, Russia
[b] Institute of Physics, Humboldt-University at Berlin, Newtonstr.15, 12489, Berlin, Germany

HIGHLIGHTS

- Correct estimating of the linear fit parameters in the presence of large outliers.
- The median of the empirical distribution of the residues determines line's shift.
- The minimum of interquantile width determines line's slope (1st method).
- The maximum of characteristic function's residues determines line's slope (2nd method).

ARTICLE INFO

Article history:
Received 16 December 2014
Received in revised form 17 March 2015
Available online 29 April 2015

Keywords:
Lévy noise
Data processing
Linear fit

ABSTRACT

We consider the problem of linear fitting of noisy data in the case of broad (say α-stable) distributions of random impacts ("noise"), which can lack even the first moment. This situation, common in statistical physics of small systems, in Earth sciences, in network science or in econophysics, does not allow for application of conventional Gaussian maximum-likelihood estimators resulting in usual least-squares fits. Such fits lead to large deviations of fitted parameters from their true values due to the presence of outliers. The approaches discussed here aim onto the minimization of the width of the distribution of residua. The corresponding width of the distribution can either be defined via the interquantile distance of the corresponding distributions or via the scale parameter in its characteristic function. The methods provide the robust regression even in the case of short samples with large outliers, and are equivalent to the normal least squares fit for the Gaussian noises. Our discussion is illustrated by numerical examples.

1. Introduction

The method of least squares linear regression (straight line fitting) has a very long history: it was invented in its simplest form by C.F. Gauß, but is still one of the most widespread and powerful approaches in data analysis. It may be used as a stand-alone tool to detect linear trends, or be incorporated into more complex analysis procedures, like Detrended Fluctuation Analysis proposed in Ref. [1], whose first step requires subtraction of linear trends from subpartitions of data. The standard variant of the method assumes the linear relation between the dependent variable y and the independent one x, and the existence of a random impacts on the outcomes of single measurements, represented by the noise ξ, so that

$$y_i = ax_i + b + \xi_i, \tag{1}$$

and is aimed onto extracting information about a and b from such noisy data. The standard method works well if the data are "compact", i.e. when the corresponding interval on the abscissa is homogeneously sampled and no large ordinate

* Corresponding author.
E-mail addresses: postnicov@gmail.com (E.B. Postnikov), igor.sokolov@physik.hu-berlin.de (I.M. Sokolov).

http://dx.doi.org/10.1016/j.physa.2015.04.025
0378-4371/

outliers are present. The method is essentially a parametric one and can be regarded as the maximum likelihood approach assuming the Gaussian distribution of independent errors. The challenges of more complicated samples originating from modern problems of experimental and computational physics and related fields have motivated works aimed to improve the accuracy of fits to extremely irregular data, i.e. the ones having outliers on the ordinate and on the abscissa (leverage points), or large errors in locating x_i, see Refs. [2,3] for the list of modern modifications. For this reason, a number of works discuss the criteria for a detection of these outliers with the following their elimination with respect to a prescribed cut-off level, and the regression of obtained "cleared" samples [4] or a choice of subintervals, where the influence of outliers could be negligible [5,6]. Another problem arises for the non-independent noises which themselves can show trends [7].

Even in the case of independent errors the problems arise if the noise possess a heavy-tailed distribution, i.e. generates large outliers. These are quite characteristic for a large variety of process in small nonequilibrium systems, network dynamics, econophysics, etc. [8]. Since these distributions may lack even the first moment, their processing, if keeping the principles of the least-square regression untouched, requires very specific methods [9,10] including repeated median regression, the consideration of a nested hierarchy block subdivisions for the analyzed sample, etc. For such cases non-parametric regression methods may be superior to the standard one.

In the present work we discuss two such approaches, the quantile regression as pioneered by Koenker and Basset [11], and the scale parameter regression based on the properties of characteristic functions. The methods are non-parametric (i.e. do not assume the specific form of the distribution) and robust (i.e. do not rely on the existence of its moments). Our numerical examples consider linear trend in presence of independent errors distributed according to Lévy stable laws. As a practical example, we consider geophysical data, namely the eastward component of the geomagnetic field measured on a moving Antarctic ice shelf, showing a linear trend from the motion and a combination of small and large scale fluctuations. Here the results of robust scale parameter regression are compared to conventional methods.

2. Linear regression

Before discussing the specific methods, let us shortly review the general idea (or, better, general ideas) of linear regression. Posing the regression problem starts from the assumption that the values of the dependent variable (observable) y_i linearly depend on x_i, but are subject to additive noise ξ_i, Eq. (1). We are looking for the way of inferring of the parameters a and b, delivering the best possible estimates $\hat{a}$ and $\hat{b}$ for these parameters. In the ideal situation (at least in the asymptotic setting when the total number of measurement points N gets large, $N \to \infty$) the method should give $\hat{a} = a$, $\hat{b} = b$. In praxis, this is usually done by the application of the least squares fit.

There are different ways to think about the least squares method.

First, we can follow the standard line of argumentation pertinent to statistical inference and make a maximal likelihood estimate for the parameters a and b assuming the distribution of ξ_i is Gaussian with zero mean and unknown dispersion,

$$p(\xi_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\xi_i^2}{2\sigma^2}\right).$$

In this case the probability density of a given realization of ξ_i is given by the product of such single-point distributions:

$$p(\xi_1, \ldots, \xi_N) = \prod_{i=1}^{N} p(\xi_i) = \left(\sqrt{2\pi}\sigma\right)^{-N} \exp\left(-\frac{\sum_{i=1}^{N} \xi_i^2}{2\sigma^2}\right).$$

Changing from ξ_i to y_i we get the corresponding density of the experimental outcomes $\{y_i\}$,

$$p(y_1, \ldots, y_N | a, b) = \left(\sqrt{2\pi}\sigma\right)^{-N} \exp\left(-\frac{\sum_{i=1}^{N} (y_i - ax_i - b)^2}{2\sigma^2}\right).$$

Considering the log-likelihood of a and b provided the data,

$$L(a, b|\{y_i\}) = \ln p(\{y_i\}|a, b) = const - \frac{\sum_{i=1}^{N} (y_i - ax_i - b)^2}{2\sigma^2}$$

and maximizing it with respect to a and b, we get the least square prescription for finding a and b by minimizing the sum of squared residues

$$R^2 = \sum_i [y_i - (ax_i + b)]^2 = \min.$$