Least-squares parameter estimation in the presence of outliers

Dmitry I. Kabanov

May 13, 2019

Least-squares algorithm is a widely used tool for estimating parameters of a functional relation generated by collected data. The basic formulation is the following. We are given data $\boldsymbol{D} = \{D_k\} = \{x_k, Y_k\},\ k = 1, \ldots, N$ with the corresponding errors $\{\epsilon_k\}$ for the observations Y_k. Besides, there is an assumption on the functional relation that maps independent variables $\{x_k\}$ to the noiseless data $\{F_k\}$: $F_k = f(\{x_k\}, \boldsymbol{X})$,where $\boldsymbol{X} \in \mathbb{R}^M$ is the vector of the model parameters. The aim is to estimate $\boldsymbol{X}$ such that the observations $\{Y_k\}$ fit the model as close as possible.

In the framework of the Bayesian data analysis, we can write that the distribution of $\boldsymbol{X}$ can be related to the distribution of the data:

$$\text{prob}(\boldsymbol{X}|\boldsymbol{D}) \propto \text{prob}(\boldsymbol{D}|\boldsymbol{X}) \times \text{prob}(\boldsymbol{D}), \tag{1}$$

which, by taking assumption that the all the data are independent of each other and that the observational noise obeys to the Gaussian distribution, reduces the problem of finding best estimate of X to finding a minimum of the function

$$L = \text{constant} - \sum_{k=1}^{N} R_k^2, \tag{2}$$

where $R_k = (F_k - Y_k)/\sigma_k$ are the residuals.

The above formulation of the least-squares algorithm is the most basic one, in which all observations $\{Y_k\}$ are treated equally. However, the following question arises: what if some of the observations are outliers, that is, errorneous measurements that do not follow the same pattern as other observations?

In this project we consider an extension of the above procedure to accomodate the presence of outliers. To conduct this, we extend Eq. (1) by adding extra parameter σ which represents that observation error over some threshold σ_0; then the likelihood function in Eq. (1) becomes the marginal PDF. Consequently, the right-hand side of Eq. (2) will change, and the new formulation will allow to estimate parameters X more robustly.

We consider two distributions of the error σ:

$$\text{prob}_1(\sigma|\sigma_0) = \sigma_0/\sigma^2 \text{ for } \sigma \geq \sigma_0 \text{ and } 0 \text{ otherwise}$$

and

$$\text{prob}_2(\sigma|\sigma_0) = \beta\,\delta(\sigma - \gamma\sigma_0) + \left(1 - \beta\right)\delta(\sigma - \sigma_0),$$

where $0 \leq \beta << 1$ and $\gamma >> 1$, and δ is the Dirac delta function.

We apply all three formulations of the least-squares procedure to synthetic data with added small-amplitude Gaussian noise, and several data points being corrupted by large-amplitude noise. To assess the performance of different formulations, we compare obtained parameter estimates and their uncertainties.