Dimension Reduction and Nonparametric Regression: A Robust Combination

C. Becker[1]

[1] Department of Statistics, University of Dortmund, D-44221 Dortmund, Germany

Keywords: DAME, Robustness, SIR, Taut String Method.

1 Introduction

In modern statistical analysis, we often aim at determining a functional relationship between some response and a high-dimensional predictor variable. It is well-known that estimating this relationship from the data in a nonparametric setting can fail due to the curse of dimensionality. But a lower dimensional regressor space may be sufficient to describe the relationship of interest.

In the following, we consider the two main steps of a combined procedure in this setting: the dimension reduction step and the step of estimating the functional relation in the reduced space. The occurrence of outliers can disturb this process in several ways. When finding the reduced regressor space, the dimension may be wrongly determined. If the dimension is correctly estimated, the space itself may not be found correctly. As a consequence, it may happen that the functional relationship cannot be found, or an incorrect relation is determined. If both, dimension and space are correctly identified, outliers may still influence the function estimation. Hence, we aim at constructing robust methods which are able to detect irregularities such as outliers in the data and at the same time can adjust the dimension and estimate the function without being affected by such phenomena.

2 The Method

Consider the situation that, given some response variable $Y \in \mathbb{R}$ and explanatory variables $X_1, \ldots, X_d \in \mathbb{R}$, we assume a functional relationship of the form $Y = g(X_1, \ldots, X_d, \varepsilon) = g(\boldsymbol{X}, \varepsilon)$, where $\boldsymbol{X} = (X_1, \ldots, X_d)^T$ is some $\mathbb{R}^d$-valued random vector, $E(\boldsymbol{X}) = \boldsymbol{\mu} \in \mathbb{R}^d$, $Cov(\boldsymbol{X}) = \boldsymbol{\Sigma}$ positive definite and symmetric, $\boldsymbol{X}, \varepsilon$ are stochastically independent. The link function g is unknown and the aim is to estimate g in a suitable way, based on a sample $(y_i, \boldsymbol{x}_i)$ of size n, $y_i \in \mathbb{R}, \boldsymbol{x}_i \in \mathbb{R}^d$, $i = 1, \ldots, n$. However, if the dimension d of $\boldsymbol{X}$ is too large, then the well-known curse of dimensionality lets usual nonparametric regression methods fail in such a case (Friedman, 1994).

One possibility to deal with this problem is to find out whether the dimension of the regressor space can be reduced in such a way that the reduced space still contains the important information on the relation between Y and $\boldsymbol{X}$. In this case, after estimating the reduced space, estimation of the functional relationship (now being $Y = f(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_K^T \boldsymbol{X}, \varepsilon)$, where $\boldsymbol{\beta}_i \in \mathbb{R}^d$, $i = 1, \ldots, K$, are unknown so-called dimension reducing directions, and $K \ll d$) can be done within this lower dimensional space. A method to estimate the dimension reduced regressor space in the above mentioned regression setting is Sliced Inverse Regression (SIR, Li, 1991). Gather et al. (2002) show that SIR may be very prone to outliers in the regressor variable $\boldsymbol{X}$. They introduce a robustified dimension adjustment method (DAME, Gather et al., 2001), referring to Li's (1991) fundamental approach, but replacing all classical nonrobust estimators used in SIR by robust ones.

It is near at hand to combine DAME with a procedure for outlier detection. We cannot do the same with SIR because the estimators used therein are themselves rather susceptible for the influence of outliers. Outliers in Y do not affect SIR but may cause trouble in the next step. Hence, the estimation of f should also be done by methods which are insensitive against outliers.

As pointed out for example by Cook (1998), in practical situations we often find that the dimension K of the reduced space equals one. We therefore and for didactic reasons restrict ourselves to this case here. A nonparametric method for estimating f in the case of a univariate regressor variable is the so-called taut strings method (Davies and Kovac, 2001). The idea is to determine the number and locations of the extremes of f and to approximate f by a piecewise constant function with an according number and according locations of extremes. To get rid of the step function, after applying the run method we can of course smoothe the resulting curve.

References

R.D. Cook (1998). Principal Hessian Directions Revisited. *Journal of the American Statistical Association*, 93, 84–100.

L. Davies, and A. Kovac (2001). Local Extremes, Runs, Strings and Multiresolution (with discussion and rejoinder). *Annals of Statistics*, 29, 1–65.

J.H. Friedman (1994). An Overview of Predictive Learning and Function Approximation. In: V. Cherkassky et al., editors, *From Statistics to Neural Networks. Theory and Pattern Recognition Applications*, pp. 1–61, Springer: Berlin.

U. Gather, T. Hilker, and C. Becker (2001). A Robustified Version of Sliced Inverse Regression. In: L.T. Fernholz et al., editors, *Statistics in Genetics and in the Environmental Sciences*, pp. 147–157, Birkhäuser: Basel.

U. Gather, T. Hilker, and C. Becker (2002). A Note on Oulier Sensitivity of Sliced Inverse Regression. To appear in *Statistics*.

K.-C. Li (1991). Sliced Inverse Regression for Dimension Reduction (with discussion). *Journal of the American Statistical Association*, 86, 316–342.

Please fill in this form and mail it together with your abstract.

My abstract fits best to topic number ... (**“PLEASE INDICATE HERE THE NUMBER FROM THE LIST OF TOPICS BELOW WHICH BEST FITS TO YOUR ABSTRACT”**.).

List of Topics:

1. Algorithms
2. Applications
3. Biostatistics
4. Computing and graphics
5. Data analysis
6. Data mining
7. Economics, finance
8. Efficiency and robustness
9. Functionals and bias
10. Fuzzy statistics
11. Geostatistics
12. Inference for robust methods, model testing
13. Location depth and regression depth

xxx 14. Multivariate methods xxx

15. Neural networks
16. Rank-based methods
17. Regression quantiles, trimming
18. Robust covariance
19. Robust designs
20. Robust regression
21. Time series analysis
22. Wavelets
23. Other (please specify)