Dimension Adjustment Methods

Claudia Becker
University of Dortmund, Department of Statistics
Vogelpothsweg 87
D-44221 Dortmund, Germany
cbecker@statistik.uni-dortmund.de

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. Determining the optimal reduced regressor dimension and the corresponding space is a pre-step of exploratory nature in the procedure of estimating a nonparametric link function in a high-dimensional setting. To complete the procedure, the final step of estimating a link function after the dimension reduction must also be taken into account. The whole procedure is considered here under the aspect of its robustness against outliers.

2. Basic Model and Notation

A situation often considered in nonparametric regression is that, given some response variable $Y \in \mathbb{R}$ and explanatory variables $X_1, \ldots, X_d \in \mathbb{R}$, a functional relationship of the form $Y = g(X_1, \ldots, X_d, \varepsilon) = g(\boldsymbol{X}, \varepsilon)$ is assumed, 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 with a sample of reasonable size, it is not possible to fill the regressor space densely enough with observations. This 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 can be done within this lower dimensional space. As proposed by Li (1991), we assume that, instead of $Y = g(\boldsymbol{X}, \varepsilon)$, there exists only a relationship between Y and a number K of linear combinations of the X_i, $i = 1, \ldots, d$:

$$Y = f(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_K^T \boldsymbol{X}, \varepsilon),$$

where $\boldsymbol{\beta}_i \in \mathbb{R}^d$, $i = 1, \ldots, K$, are unknown so-called dimension reducing directions, and $K \ll d$.

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 f 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 estimation of f. 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 f without being affected by such phenomena.

3. Robust dimension reduction

A method to estimate the dimension reduced regressor space in the above mentioned regression setting is Sliced Inverse Regression (SIR, Li, 1991). The idea behind this method is that under suitable design conditions, the appropriately normalized inverse regression curve (i.e. the conditional expectation of $\boldsymbol{X}$ given Y) lies almost surely in the linear subspace spanned by the directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$. The inverse regression function is very roughly estimated by a vector valued "step function". The K directions of maximal variability of these vectors (gained by a principal component analysis) yield the estimates for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$.

Many dimension reduction methods suffer from a certain sensitivity against the influence of outlying observations (e.g. principal component analysis, Croux and Haesbroeck, 2000; factor analysis, Pison et al., 2000). Gather et al. (2001b) 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., 2001a), 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. In the first step of DAME, analogously to the first step of SIR, the $\boldsymbol{X}$ data are standardized w.r.t. location and covariance. In contrast to SIR, DAME uses certain robust estimators which can also be used for simultaneous outlier detection (Becker and Gather, 1999). 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.

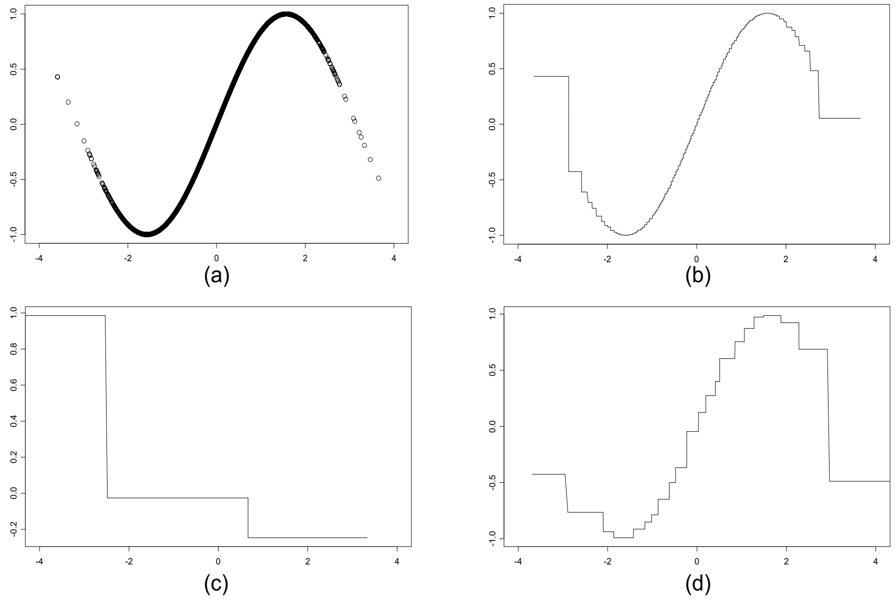
4. Robust nonparametric regression

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 run method (Davies, 1995). 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. This function additionally has to satisfy a certain "run condition", meaning that the signs of the residuals of the approximation do not form long sequences (runs). It can be interpreted as "residuals looking like white noise". The run method can cope with data situations where outliers in Y occur (especially blocks of outliers or large proportions of isolated outliers, cf. Davies and Kovac (2001)) while still behaving quite well if there are no outliers at all. To get rid of the step function, after applying the run method we can of course smoothe the resulting curve.

5. Example

We consider a simple example to illustrate the application of the complete process of dimension reduction and regression estimation with the procedures described above. We take a data set of $\boldsymbol{X}$ observations of size $n = 5000$ in 9 dimensions generated according to $\boldsymbol{X} \sim N(\mathbf{0}, \mathbf{I})$, and let $Y = \sin(X_1)$. Hence, $K = 1$ and $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$. Then we construct an outlier in direction X_1 by setting one element of X_1 to 10000. The results of applying both, SIR and DAME, in combination with the run method to the data are shown in Figure 1. It is obvious that the combination SIR/RUN works quite well in the case of using the original data (without

Figure 1. True relationship and estimated functions



(a) undisturbed data, (b) estimation with SIR/RUN, undisturbed data, (c) estimation with SIR/RUN, data with one outlier in $\boldsymbol{X}$, (d) estimation with DAME/RUN, data with one outlier in$\boldsymbol{X}$

the outlier) but fails completely when taking the disturbed data. In contrast to this, the robust combination DAME/RUN yields far better results.

REFERENCES

Becker, C., Gather, U. (1999). The masking breakdown point of multivariate outlier identification rules. *J. Amer. Statist. Assoc.*, **94**, 947–955.

Cook, R.D (1998). Principal Hessian Directions Revisited. *J. Amer. Statist. Assoc.*, **93**, 84–100.

Croux, C., Haesbroeck, G. (2000). Principal Component Analysis Based on Robust Estimators of the Covariance or Correlation Matrix: Influence Functions and Efficiencies. *Biometrika*, **87**, 603–618.

Davies, P.L. (1995). Data Features. *Statist. Neerlandica*, **49**, 85–245.

Davies, P.L., Kovac, A. (2001). Modality, runs, strings and multiresolution. *To appear in Ann. Statist.*

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

Gather, U., Hilker, T., Becker, C. (2001a). A Robustified Version of Sliced Inverse Regression. *To appear in Proc. of the Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics, Ascona, May 23-28, 1999.*

Gather, U., Hilker, T., Becker, C. (2001b). A Note on Outlier Sensitivity of Sliced Inverse Regression. *Preprint.*

Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction (with discussion). *J. Amer. Statist. Assoc.*, **86**, 316–342.

Pison, G., Rousseeuw, P.J., Filzmoser, P., Croux, C. (2000). A Robust Version of Principal Factor Analysis. In: Bethlehem, J.G., van der Heijden, P.G.M. (eds.), *COMPSTAT 2000. Proceedings in computational statistics.* Physica, Heidelberg, 385–390.

RESUME

Nous considérons le procédé qui permet d'estimer une relation fonctionnelle entre une réponse univariée et un prédicteur de grande dimension. Nous étudions une approche en deux étapes qui associe la réduction de la dimension de la variable de régression avec l'estimation de la fonction dans l'espace réduit. Le but principal est la robustesse d'une telle méthode par rapport à la présence de valeurs aberrantes dans les données.