The Curse of Dimensionality – A Challenge for Mathematical Statistics

U. Gather, C. Becker, Dortmund

Mathematics Subject Classification: 62-02, 62-07, 62H99

Keywords and Phrases: curse of dimensionality, dimension reduction, high-dimensional data, multivariate methods.

1 Introduction

It is a well-known fact that in almost any field of applied research, nowadays progress is achieved mainly or to a large extent by analyzing and properly interpreting either experimental or empirical data. One may think of the life sciences here (e.g. analyzing the human genome, understanding the functioning of the brain) or of the engineering sciences (e.g. understanding the influence of process parameters on the final quality properties of a product).

For at least two reasons the data which have to be analyzed here to gain more information are very high-dimensional. The first reason is that the open research questions themselves have become more and more complex. Already known relationships – e.g. in medicine – can only be described by functions between high-dimensional influential and response variables. The second reason comes from the immense increase of facilities to collect, store and process high-dimensional and massive data by modern computer power, which makes it possible for the first time to actually investigate such complex research questions. In any case, the claim for statistical methods and data analytical tools which are able and appropriate to extract the essential information from high-dimensional data structures is a real and urgent one.

From a pure mathematical first perspective one might argue that statistical methods for multivariate data do already exist, and very often their (in some sense) optimal performance has been proved, independently of the dimension of the data. However, optimal can be very bad in high dimensions.

It is the objective of the first part of this paper to show the principal and algorithmic problems which constitute the new challenge for mathematical statistics here. In the second part, some procedures will be presented which are constructed to cope with the curse of dimensionality. These will be prototypical examples of approaches, for which we will also investigate some new problems they are still generating, leaving us with enough open questions in mathematical statistics.

The oral presentation of this paper was given as a plenary lecture at the Annual Conference of the DMV (German National Mathematical Society). It was meant to give a partial overview on the problem of analyzing high-dimensional data and was addressed to an audience with a general mathematical background not necessarily familiar with special methods in mathematical statistics and data analysis.

We therefore chose to start at a rather elementary level and to show principal problems as well as some solutions by means of prototypical examples such as regression modeling with high-dimensional influential variables and outlier identification in high-dimensional data from simple distributions.

The structure of this paper is as follows. In Section 2, we repeat some well-known facts about multidimensional data and explain what we mean by the curse of dimensionality. In Section 3, we consider some very simple statistical models to in order to motivate the need for new methods in high dimensions. Finally, two special statistical tasks are discussed in more detail in the situation of high-dimensional data. These are the identification of outliers and dimension reduction in nonparametric regression. We finish with some concluding remarks.

2 The curse of dimensionality: facts, examples, real life data

In this section, we first restrict ourselves to examples with three-dimensional data in order to be able to visualize structures. It can be seen that typical difficulties already begin in this seemingly simple situation.

Example 2.1. *Let $\mathcal{X}_a$, $\mathcal{X}_b$ be the sets of two nested cylinders of infinite height in $\mathbb{R}^3$, and consider random variables $\boldsymbol{X}_a$, $\boldsymbol{X}_b$ with values in $\mathcal{X}_a$, $\mathcal{X}_b$, respectively, i.e. $\boldsymbol{X}_a = (X_{a1}, X_{a2}, X_{a3})^T$, $\boldsymbol{X}_b = (X_{b1}, X_{b2}, X_{b3})^T$. For $\boldsymbol{X}_a$, the bivariate marginal distribution of $(X_{a1}, X_{a2})^T$ is assumed as uniform on the boundary of a two-dimensional circle with radius 0.5, centered at the origin,*

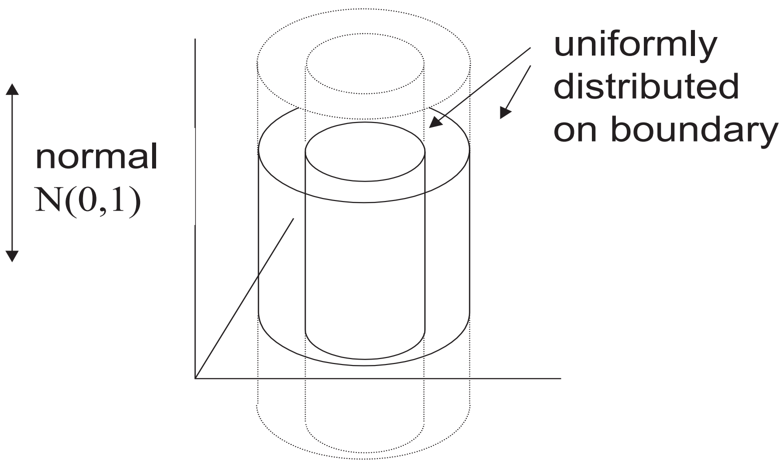


Figure 1: Nested cylinders

while X_{a3} *is standard normally distributed on* $\mathbb{R}$. *For* $\boldsymbol{X}_b$ *we assume a similar structure, but the bivariate marginal is uniform now on the boundary of a circle with radius 1 (see Figure 1). We take a sample of size 500, where we have 250 observations of* $\boldsymbol{X}_a$, $\boldsymbol{X}_b$ *each, and try to re-identify the given structure. Figure 2 shows all possible one- and two-dimensional projections of the data onto the coordinate planes and axes. Obviously the structure is visible in some of the two-dimensional projections, but we also see that indeed only* **some** *of the projections are suitable. In the one-dimensional projections which only give us the empirical marginal distributions, we cannot detect any structure.*

While in Example 2.1 we can still identify some of the structure by projecting into lower dimensions, the following example shows a situation where the lower dimensional projections do no longer reveal any information.

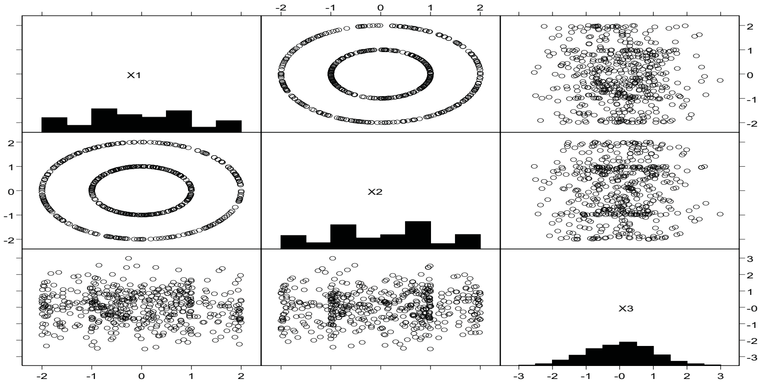


Figure 2: Nested cylinders data: one- and two-dimensional projections

Example 2.2. *Let $\mathcal{X}_a$, $\mathcal{X}_b$ be the sets of two nested spheres in $\mathbb{R}^3$, and consider again random variables $\boldsymbol{X}_a = (X_{a1}, X_{a2}, X_{a3})^T$, $\boldsymbol{X}_b = (X_{b1}, X_{b2}, X_{b3})^T$ with values in $\mathcal{X}_a$, $\mathcal{X}_b$, respectively. Now, $\boldsymbol{X}_a$ and $\boldsymbol{X}_b$ are uniformly distributed on the boundaries of three-dimensional spheres, centered at the origin, with radii 0.5 and 1, respectively. Figure 3 indicates the structure. As in Example 2.1, we take a sample of size 500, where we now choose 150 observations of $\boldsymbol{X}_a$ and 350 observations of $\boldsymbol{X}_b$. The two- and three-dimensional projections of these data onto the coordinate planes and axes can be seen in Figure 4. This time, the structure is no longer visible in any of the projected data sets. One just might suspect that there is some inherent structure.*

From the very simple Examples 2.1 and 2.2 it becomes clear that already for three-dimensional data, it can become difficult to detect existing structures by means of lower dimensional projections. On the other hand, in high-dimensional data situations, where we also lack geometrical imagination,

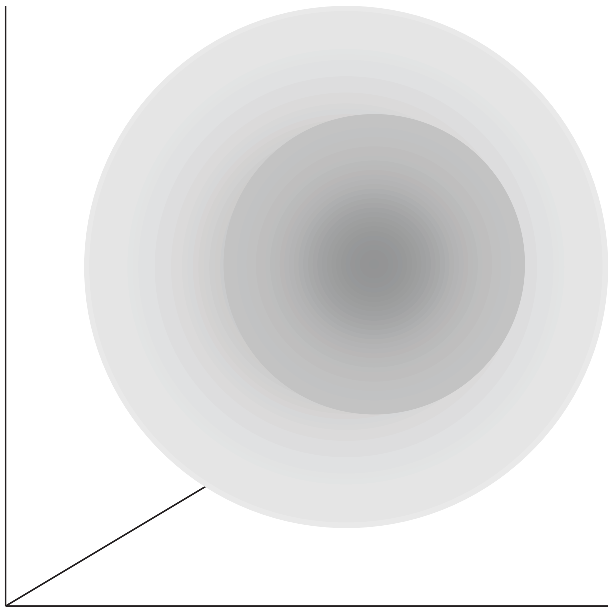


Figure 3: Nested spheres

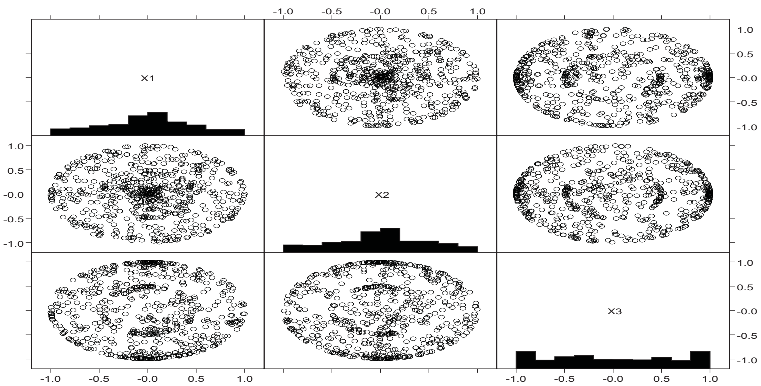


Figure 4: Nested spheres data: one- and two-dimensional projections

the detection of structures heavily depends on such projections. Hence, generally we do not have any chance to gain informative visualizations by simply projecting high-dimensional data into lower-dimensional spaces.

There is another problem coming along with high dimensions, and this is a principal one which we will illustrate by an example from nonparametric regression.

Example 2.3. *Consider a univariate response variable $Y \in \mathbb{R}$ which depends on a multivariate random regressor variable $\boldsymbol{X} \in \mathbb{R}^d$ by some unknown link function g, $Y = g(\boldsymbol{X}, \varepsilon)$, where ε denotes an additional error term. Standard methods to estimate g in such a nonparametric regression setting use nearest neighbour techniques. That means, to estimate $g(\boldsymbol{x}_0)$ at some point $\boldsymbol{x}_0 \in \mathbb{R}^d$, we take observations of $\boldsymbol{X}$ and the corresponding Y values in a neighbourhood of $\boldsymbol{x}_o$. But this method fails in high dimensions. If, for example, we want to estimate the value of g at the origin, we have to face the fact that even in a quite large neighbourhood of the origin, in a ball with radius 0.5, say, we find observations of $\boldsymbol{X}$ only with very small probability (see Table 1). Even worse: with increasing dimension d, this probability rapidly decreases towards zero.*

The conclusion drawn from Example 2.3 is that even for very large samples in $\mathbb{R}^d$, say from $d = 5$, the sample space is only sparsely filled and all observations are rather far away from each other. Consequently, for estimation purposes we must either take into account points which are very distant – yielding a large bias of an estimator – or we are confronted with the problem of not

Table 1: Probability of $\boldsymbol{X}$ to lie in a sphere with radius 0.5 around the origin for d-dimensional uniformly and normally distributed $\boldsymbol{X}$

	dimension d			
$P(\lVert\boldsymbol{X}\rVert \leq 0.5)$	1	5	10	20
$\boldsymbol{X} \sim U_d[0,1]$	0.5	0.0313	0.001	$< 10^{-6}$
$\boldsymbol{X} \sim N_d(0,1)$	0.3839	0.1175	$< 10^{-6}$	≈ 0

taking enough observations – yielding a large variance of the estimator.

This phenonemon that the high-dimensional sample space is hardly filled densely enough with observations even for massive data sets, is called the "curse of dimensionality" (Bellman, 1961, Friedman, 1994).

There is a further well-known problem connected to the curse of dimensionality, which is an algorithmic one. We illustrate this problem by the following example which also discusses a new method in robust regression.

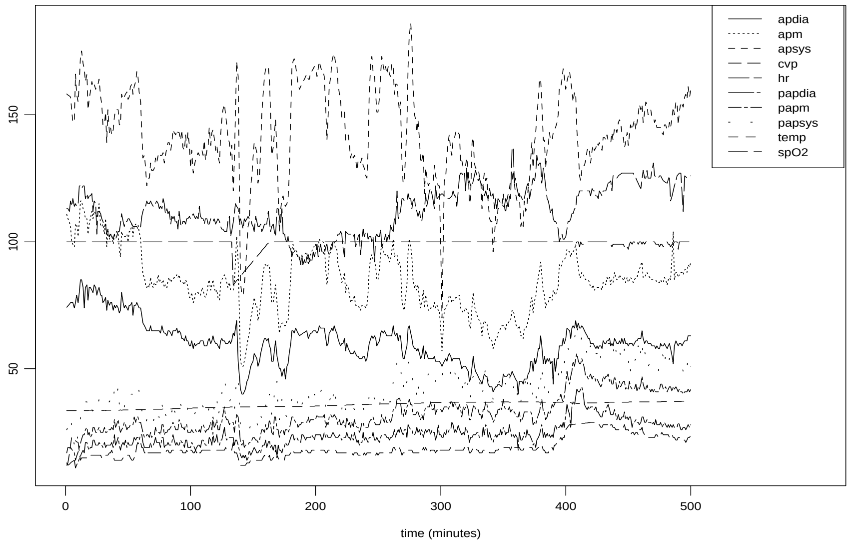
Example 2.4. *Consider a data set* $(x_{1,i}, \ldots, x_{d,i}, y_i) \in \mathbb{R}^{d+1}, i = 1, \ldots, n,$ *consisting of observations of a one-dimensional dependent variable* Y *and a d-dimensional regressor* $\boldsymbol{X} = (X_1, \ldots, X_d)^T$. *The idea of Rousseeuw and Hubert (1999) is to look for a best linear fit, not with respect to least squares, but as a hyperplane which lies in some sense "as deep as possible" in the data. This leads to the definition of the so-called* **regression depth** *which can be seen as a generalization of the median to the regression setting. For a formal definition see*

Rousseeuw and Hubert (1999). The method is highly robust against deviations from the assumption of a normal distribution for $\boldsymbol{X}$ *(for the concept of robustness cf. Huber, 1981). It also has caught attention in analytical geometry and within complexity theory. By use of methods from these fields it has been shown that every algorithm for determining the deepest regression fit exactly must be of order* $\mathcal{O}(n^{2d+1}\log(n))$ *for* $d > 1$ *(Amenta et al., 2000). This means that the runtime of every exact algorithm grows exponentially with the dimension* d.

Altogether, we can summarize the difficulties for the statistical analysis of high-dimensional data as:

- the problem of visualization,
- the curse of dimensionality,
- the numerical and algorithmic realization of valuable methods.

The problem of analyzing high-dimensional data is of course not only interesting from a theoretical point of view. Such data sets occur in real-world problems. We illustrate this by two brief examples. The first deals with data from a biopsychological experiment. For a sample of test persons, certain hormone levels and several intellectual as well as physiological abilities were observed. The goal was to find out about a possible interrelationship between hormone levels and abilities to solve certain tasks (cf. Hausmann et al., 2001), and the related study dealt with a total of 30 variables, yielding indeed a high-dimensional situation.



Variables:

systolic arterial pressure (apsys), diastolic arterial pressure (apdia),

mean arterial pressure (apm), temperature (temp),

central venous pressure (cvp), heart rate (hr),

systolic pulmonary artery pressure (papsys),

diastolic pulmonary artery pressure (papdia),

mean pulmonary artery pressure (papm), pulsoximetry (spO2)

Figure 5: Data from intensive care medicine

The second example comes from intensive care medicine, where one has to analyze online-monitoring data (Bauer et al., 1998). Figure 5 shows a small part of the vital signs of a critically ill patient in the course of time, which were recorded online at the bedside. Nowadays, clinical information systems constantly record up to 2000 variables, some of them in very short time intervals. The complex dependence structure of the various signals is obvious, and additionally there also exists a strong dependence in time. The long-term objective here is online modeling and online analysis of such series for some kind of decision support. It is easily seen that this objective is considerably obstructed by the curse of dimensionality.

What does statistics have to offer in such high-dimensional data situations? In the following section, we explain and discuss some modern approaches for dealing with these situations.

3 Models and methods

We concentrate on two simple statistical models here and start with the case of a d-dimensional random variable $\boldsymbol{X}$ distributed according to F, an element of some given class $\mathcal{F}$ of distributions. Typical tasks in this situation are to estimate certain features of F such as the expectation $E(F)$ or the covariance matrix $Cov(F)$, or to find "interesting structures" in a sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ of $\boldsymbol{X}$, i.e. to learn more about the unknown F. Statistical methods for this latter purpose are often called "data mining" procedures (Hand, 1998). Methods in this model class include

- the grand tour (Asimov, 1985) and other projection pursuit methods (Huber, 1985),
- methods for dimension reduction like
  - principal component analysis (cf. Jolliffe, 1986, Pearson, 1901),
  - factor analysis (cf. Lawley and Maxwell, 1971, Pison et al., 2000),
  - neural networks (cf. White, 1989),
- robust covariance estimation (e.g. for outlier detection, compare Davies, 1987, Rocke, 1996, Rousseeuw, 1985).

Examples 2.1 and 2.2 already showed that projection based procedures may not always really be helpful in finding structures. However, projection pursuit methods are quite popular since they are sometimes the only possible choice and in their sophisticated versions they might be able to render a first clue of the data generating mechanism. We will therefore give a short view on the projection pursuit principle in the next section.

As a second model we will consider the simple high-dimensional nonparametric regression model from Example 2.3. Key methods in this situation are

- neural networks (cf. White, 1989),
- the support vector machine (cf. Christmann et al., 2000, Vapnik, 1998),

- projection pursuit regression (Friedman and Stuetzle, 1981),
- MARS / MART (Friedman, 1991, 1999a,b),
- dimension reduction methods like SIR (Li, 1991), PHD, SAVE (Cook, 1998, Cook and Weisberg, 1991, Li, 1992), DAME (Gather et al., 2001a).

We will briefly comment on neural networks in this context.

3.1 Projection pursuit methods

The so-called grand tour (Asimov, 1985) consists of the automatic animation of two-dimensional projections of a d-dimensional data cloud. This results in a three-dimensional rotation of the projections and mainly is a tool for visualization of high-dimensional data. As we have seen in Examples 2.1 and 2.2, such a method might not be helpful in finding structures in a data set. Moreover, the grand tour will discover interesting projections (if such exist) only in passing by, as it does not search for those systematically. In contrast to this, projection pursuit methods (Huber, 1985) do a directed search for "interesting" projections. Their focus can be e.g. to detect a nonlinear structure, holes in the point cloud, the central mass, or main directions of variability.

The directed search in such projection pursuit methods is performed by means of maximizing (or minimizing) a certain real-valued projection index I. Let $\boldsymbol{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ be a sample of size n of some d-dimensional random variable $\boldsymbol{X}$, then a projection pursuit method determines that linear projection

A for which $I(A\boldsymbol{X}_n) = \max$, where the maximum is taken over all possible linear projections into $\mathbb{R}^p, p < d$, and $I(A\boldsymbol{X}_n) = I((A\boldsymbol{x}_1, \ldots, A\boldsymbol{x}_n))$. The value of $I(A\boldsymbol{X}_n)$ can be interpreted as the degree of interest of the projection. As the maximization takes place in $\mathbb{R}$, and since we project into some lower-dimensional space, we thus escape from the curse of dimensionality. A very simple well-known example of a projection pursuit method is principal component analysis (e.g. Jolliffe, 1986), where we project into $\mathbb{R}$, and the projection index equals the sample variance, $I(A\boldsymbol{X}_n) = Var(A\boldsymbol{X}_n)$.

Compared to the grand tour, projection pursuit methods can be seen as more advanced. But the search is computationally rather expensive, and the problems of visualization as shown before still occur.

3.2 Neural networks for regression

The general idea of neural network methods is to copy the idea of efficient information management of biological neural systems by clever mathematical modeling. This is done by assuming that certain input variables yield a certain output by hidden links (see Figure 6 for a schematic description). Neural networks are used in various applications to approximate unknown complex functional relationships.

To forecast the output from given observations of the input variables, the network has to be "trained" first. Translated into statistical terms, this means we have to estimate the network's parameters. These are part of the

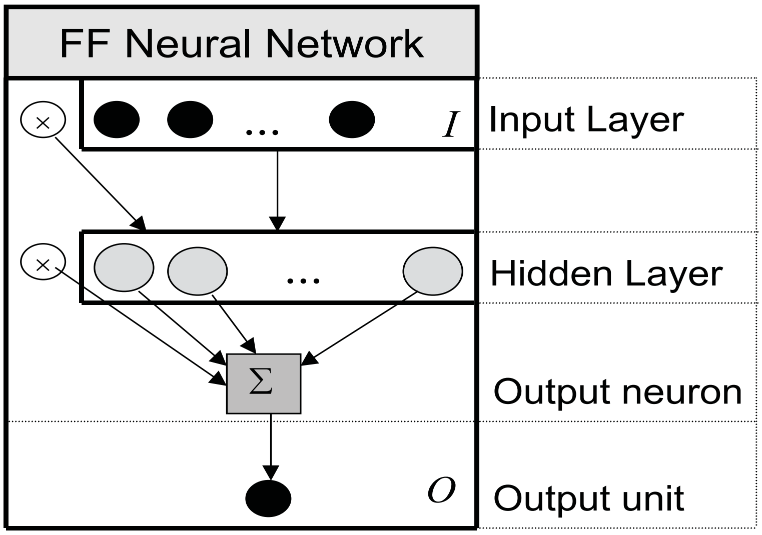


Figure 6: Feedforward neural network, schematic description

mathematical-statistical model for a feedforward neural network, given in the following form:

$$y = \text{output}(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{x}^T \boldsymbol{\alpha} + \sum_{j=1}^{M} \beta_j g_j \left(\sum_{i=1}^{d} \gamma_{ji} x_i \right),$$

where $\boldsymbol{x} = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$ denotes an input variable, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)^T$ and $\boldsymbol{w} = (\beta_1, \ldots, \beta_M, \gamma_{1,1}, \ldots, \gamma_{M,d})^T$ are the vectors of network weights, and $g_j(.), j = 1, \ldots, M$, are the so-called activation functions. These functions and also their number, M, have to be chosen beforehand. The network weights are the above mentioned network's parameters which have to be estimated or "learned" (in the neural network terminology) from the data. This is done by minimization of certain criteria. We note that – statistically seen – finding such a feedforward neural network reduces to fitting a generalized linear model to a data set of input and output observations (also see Warner and Misra, 1996). The frequently praised flexibility of the neural network approach originates from the possibility of choosing the functions g_j as desired. Also, a dimension reduction takes place in the sense that only a single index model is considered, which means that we do not model any interactions. Moreover, modern algorithms in neural network methods are numerically very stable and indeed rather flexible.

4 Some special problems in high dimensions

We now focus on some special tasks in the high-dimensional data situation. First, we consider again the simple model of a sample of size n from some d-dimensional elliptically contoured null distribution like the multivariate normal. We assume that the majority of the data in the sample come from this null distribution. Some of the observations are assumed to be outlying in the sense that they lie in a region of low probability mass of the null distribution. They are outliers in the sense of Davies and Gather (1993). We consider the task of detecting all such outliers in the sample with respect to the (at least partially) unknown target distribution.

4.1 Covariance estimation / outlier detection

We consider simultaneous rules for outlier identification, i.e. rules which detect all outliers in a sample in one step. An observation $\boldsymbol{x}_i \in \mathbb{R}^d$ is declared as outlying iff its generalized Mahalanobis distance $d_i^2 = (\boldsymbol{x}_i - \boldsymbol{m})^T \boldsymbol{S}^{-1}(\boldsymbol{x}_i -$

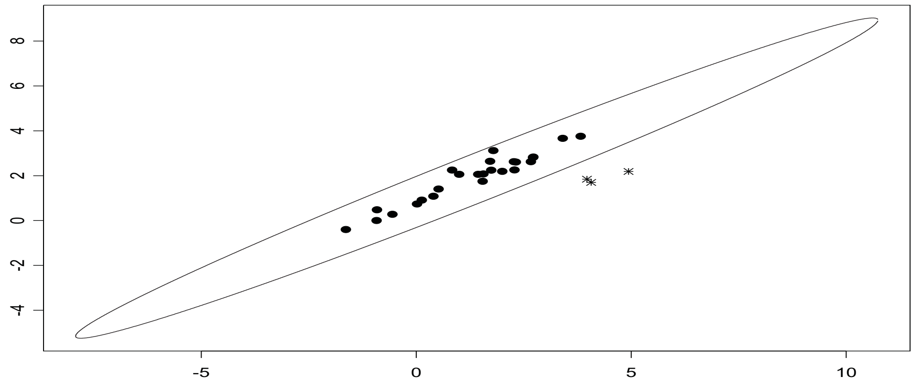


Figure 7: Simultaneous outlier identification rule: points marked as asterisks are detected as outlying

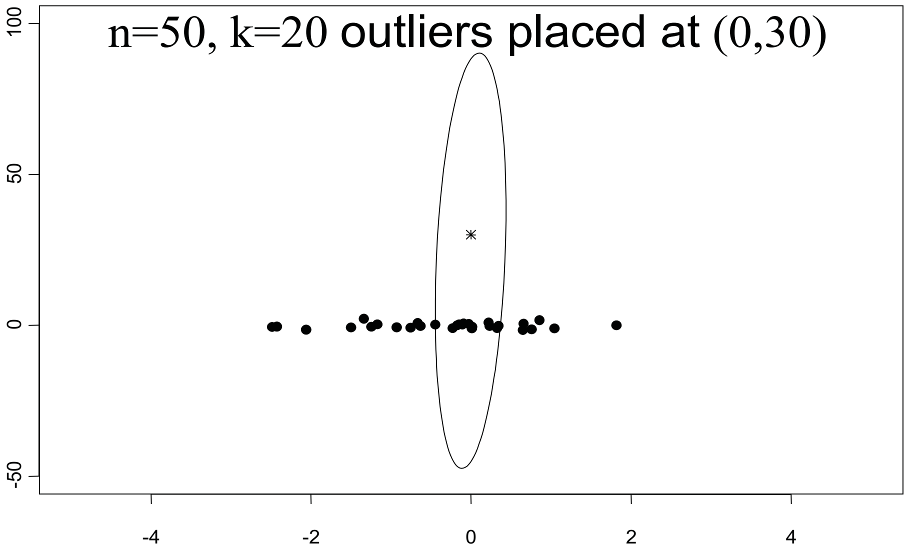
$\boldsymbol{m}$) exceeds some suitably chosen critical value $c(d, n, \alpha_n)$ (Barnett and Lewis, 1994, Becker and Gather, 1999, Rousseeuw and Leroy, 1987, Rousseeuw and van Zomeren, 1990). Here, $\boldsymbol{m}$ and $\boldsymbol{S}$ denote estimators of the d-dimensional location vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, respectively, of the underlying target distribution from which the majority of the data $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is assumed to come. The constant $c(d, n, \alpha_n)$ can for example be determined such that in a sample of size n from the d-variate normal distribution, an observation exceeds c only with probability α for some given value $\alpha \in [0, 1]$ (Becker and Gather, 1999, see also Davies and Gather, 1993). Figure 7 illustrates the idea of such an outlier identification rule. Observations lying outside the ellipsoid are identified as outlying, meaning that they lie too far away from the center of the main data cloud with respect to its shape.

Quite obviously, identification methods as just described need estimators of the center and shape of the underlying null distribution which are themselves insensitive against the influence of possible outlying observations in the sample. This means that the problem of finding the outliers in a sample is more or less equivalent to the problem of finding a "good" estimation of the covariance structure when outliers may be present. It is well known that independently of the dimension d, we need highly robust estimators of location and covariance – in the sense of maximum breakdown point – to guarantee the maximum possible protection against misclassification of observations as outliers or non-outliers (masking, swamping), respectively (Becker and Gather, 1999, Rocke, 1996). Suitable estimators are for instance the minimum volume ellipsoid (MVE) and the minimum covariance determinant (MCD) estimators (Rousseeuw, 1985), S-estimators (Davies, 1987, Rocke, 1996), as well as projection based approaches like the Stahel-Donoho estimators (Donoho, 1982, Stahel, 1981) or projection pursuit estimators (Li and Chen, 1985).

In this context, S-estimators $(\boldsymbol{m}, \boldsymbol{S})$ for location and covariance are optimal in various ways (Becker and Gather, 1999, 2001). They are defined as the simultaneous solution of the problem of minimizing the determinant of some positive definite matrix $\boldsymbol{S}$ under certain constraints concerning $\boldsymbol{m}$ and $\boldsymbol{S}$ (Davies, 1987, Rocke, 1996). Algorithmically, solutions can be found by an iterative procedure starting with highly robust initial estimates such as MVE or MCD. For details see Rocke (1996). The following example illustrates the application of a simultaneous outlier identification rule based on S-estimators.

Example 4.1. *(Becker, 2000) We take a two-dimensional sample of size* $n = 50$*, where* 30 *observations come from the bivariate standard normal, while* 20 *observations are placed at a point with coordinates* $(0, 30)$*, i.e. they are located quite distant from the true center. Figure 8 (upper part) shows this data set together with the critical ellipse for outlier identification, resulting from estimating location and covariance structure by means of a biweight S-estimator according to Rocke (1996). We can see that the estimated covariance structure does not allow for identifying the outliers, even though we use highly robust estimators. However, the procedure does not yield arbitrarily bad results as can be seen in the lower part of Figure 8. If we move the outliers to some farther distance, namely place them at the point* $(0, 50)$*, they are clearly classified as outlying by the rule. Somewhere in between* $(0, 30)$ *and* $(0, 50)$*, we find the distance at which the behaviour of the procedure changes from not identifying the outliers to detecting them. We call this distance the size of the largest nonidentifiable outlier.*

Closer investigations of the size of the largest nonidentifiable outlier can be found in Becker (2000) and Becker and Gather (2001). The results can be summarized as follows. Let us first consider the size of the largest nonidentifiable outlier as a performance criterion for a simultaneous outlier identification rule in the lower dimensional case. For samples of size $n = 20, 50$ in dimension $d = 4$, we compare the mean sizes of the largest nonidentifiable outliers found by simulation for three identification rules. The methods are based on S-estimators



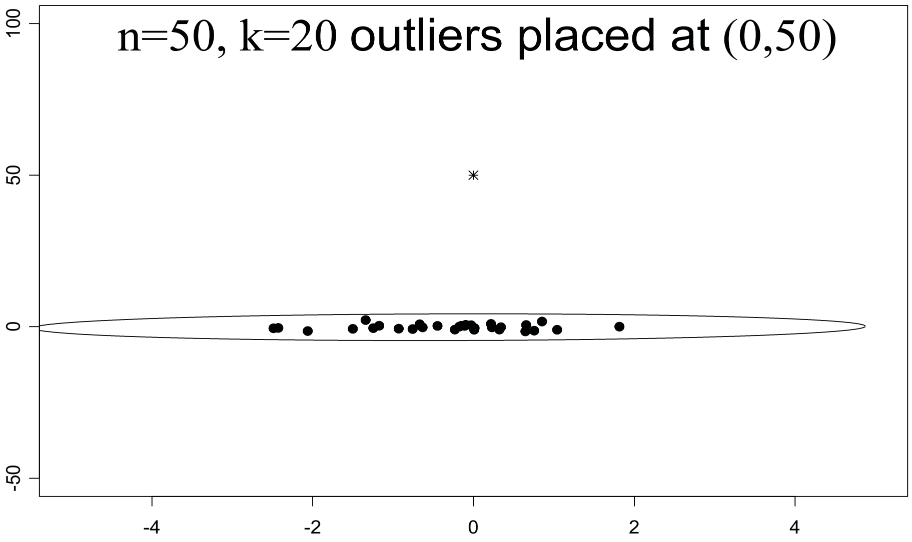


Figure 8: S-estimator identification: $n = 50$, $k = 20$ outliers placed at $(0, 30)$ and $(0, 50)$, respectively

as described before, and on the robust MVE and MCD estimators which are also used as the initial estimates in the S-estimator algorithm. Figure 9 shows the results (from Becker and Gather, 2001). We observe that

1. the solid line lies below the others, especially for large proportions of outliers in the samples; hence, the S-estimator identification yields a lower "bias" in the sense of a smaller value of the largest nonidentifiable outlier, and the S-estimator iteration obviously leads to an improved outlier identification relative to the performance of the identification based on the initial estimates;

2. there is only one line drawn for the S-estimator identification rule; for low to medium dimensions, the size of the largest nonidentifiable outlier does not depend on the choice of the initial estimate in the S-estimator iteration.

Both effects disappear in high dimensions (see Figure 10, from Becker, 2000).

1. The results are no longer independent of the choice of the initial estimate. Instead, the sizes of the largest nonidentifiable outliers are mainly identical for the MVE-identification rule and for the S-estimator identification based on an initial MVE estimate. The same is true for the MCD case.

2. This implies that the S-estimator iteration no longer provides an improvement relative to the initial estimates.

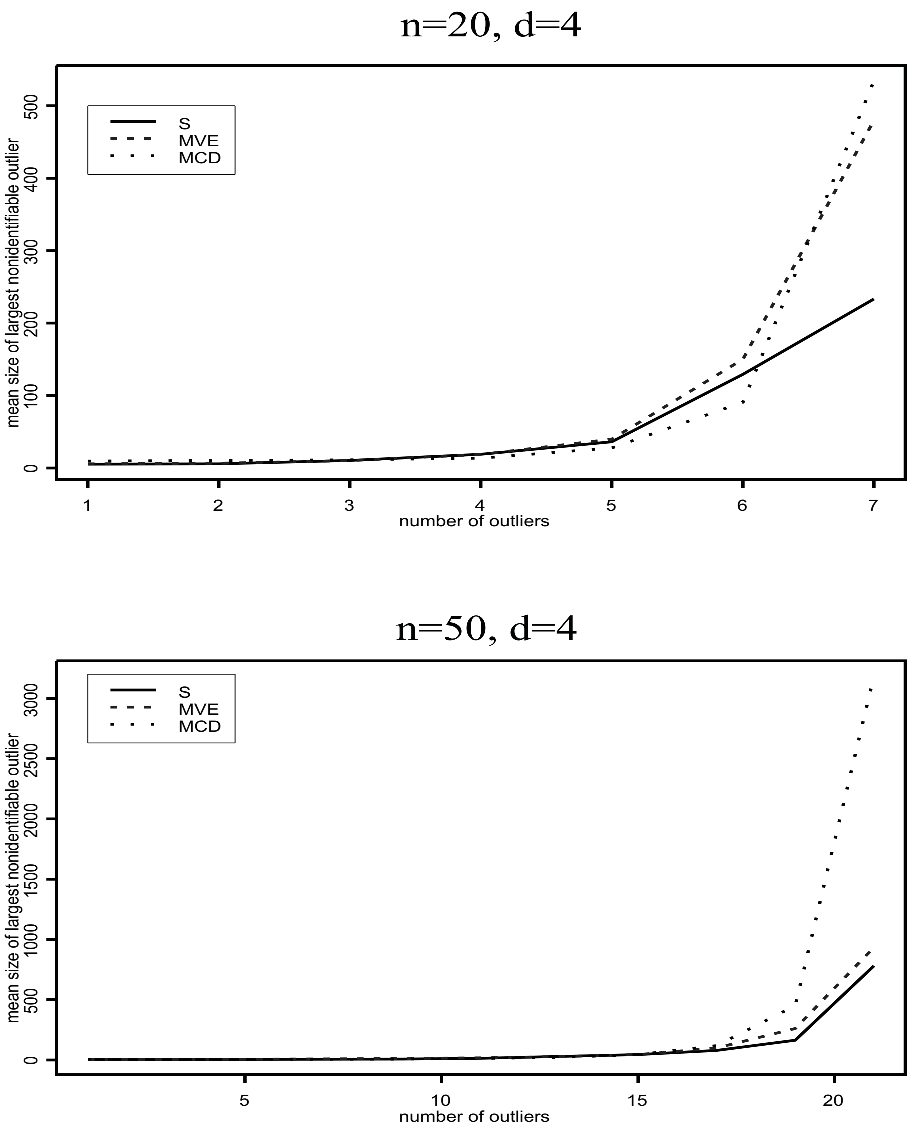


Figure 9: Mean size of the largest nonidentifiable outlier: low-dimensional case

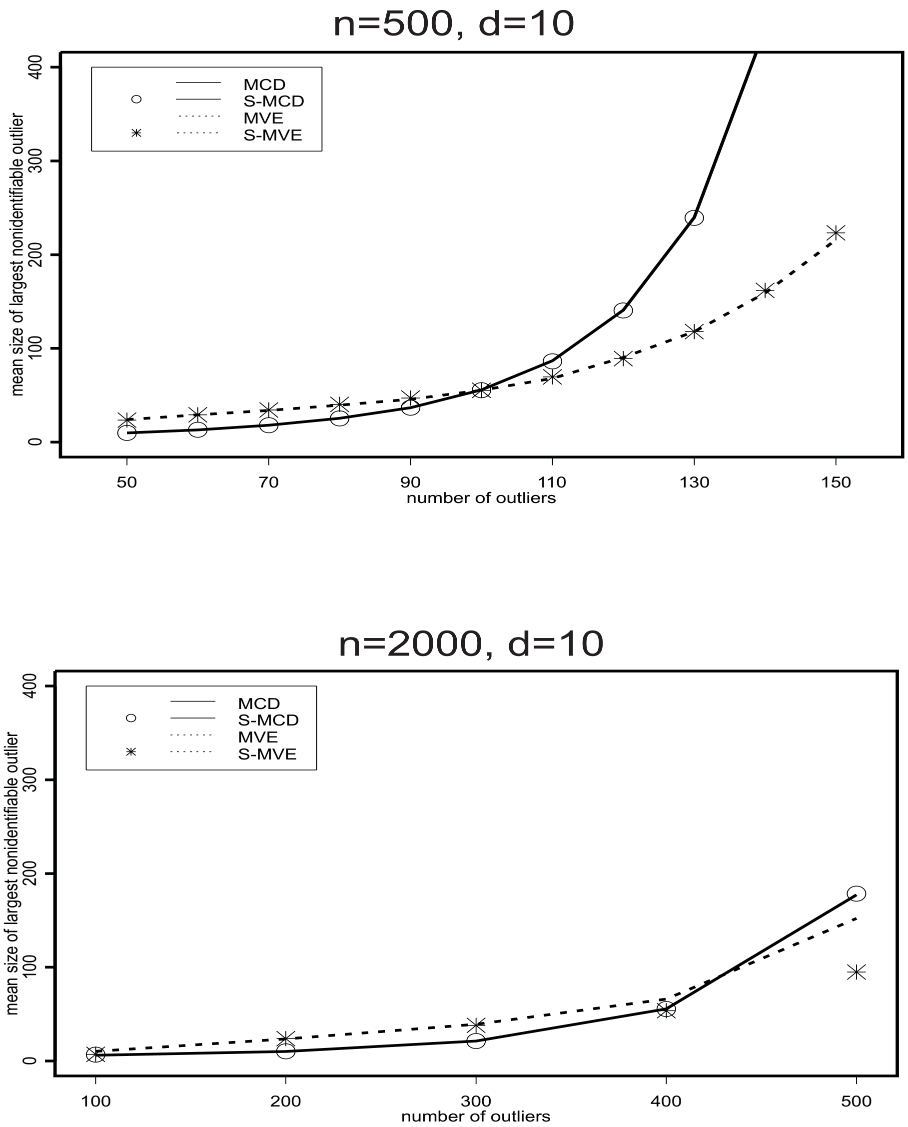


Figure 10: Mean size of the largest nonidentifiable outlier: high-dimensional case

To conclude, we find that for low to medium dimensions of the data a simultaneous outlier identification rule based on an S-estimator possesses good (theoretical) robustness properties and shows a good (practical) performance in most cases. Moreover, we often even get improved results by the S-estimator iteration, compared to the results of the rules based on the initial estimates. But for high dimensions, the well-behaved covariance S-estimator looses its advantages over the initial estimates, and the performance is no longer independent of their behaviour. Hence, we either need to improve the iterative algorithm for calculating S-estimators, or we need to construct completely new algorithms which are suitable for the use with high-dimensional data. In general, we see that there is a rather strong need for new, robust covariance estimators which are especially designed for high dimensions.

4.2 Dimension reduction in regression

In the second part of this section, we focus on the situation of a nonparametric regression model, where the regressor is d-dimensional. In Example 2.3 we have already seen why usual nonparametric regression techniques fail when d is large. A natural goal is thus to reduce the dimension of the regressor to avoid the curse of dimensionality.

To fix ideas, let us consider a univariate response $Y \in \mathbb{R}$ which depends on a d-dimensional random regressor variable $\boldsymbol{X} = (X_1, \ldots, X_d)^T \in \mathbb{R}^d$ by some unknown function $g : \mathbb{R}^d \mapsto \mathbb{R}$ and some error ε, $Y = g(\boldsymbol{X}) + \varepsilon$, $d \geq 2$. We then assume that we can write this functional relationship by means of some function $f : \mathbb{R}^{K+1} \mapsto \mathbb{R}$ (also unknown) which is defined on some lower dimensional subspace of dimension $K < d$. More exactly, the high-dimensional regressor $\boldsymbol{X}$ is projected into this subspace by so-called dimension reducing (dr) directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ such that one may write $Y = f(\boldsymbol{\beta}_1^T \boldsymbol{X}, \ldots, \boldsymbol{\beta}_K^T \boldsymbol{X}, \varepsilon)$. Hence, we assume a multi-index model, for which it seems to be reasonable to estimate the space $\mathcal{B}$ spanned by the directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ in a first step and to find the function f in the reduced space afterwards by appropriate standard methods. Sliced inverse regression (SIR; Li, 1991) is a compound procedure to estimate the dimension reducing space $\mathcal{B}$ as well as the dimension K of this space. 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 that linear subspace which is spanned by the dr directions. This fact is used for estimating the dr directions by first approximating the inverse regression function very roughly by a vector valued "step function" where the sets of Y-values with equal step function values are called "slices". Then the K directions of maximal variability of these vectors are gained by a principal component analysis. These directions yield the estimates for the dr directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$.

SIR and its modifications (Cook and Weisberg, 1991, Li, 1991, 1992, Cook, 1998) got a lot of attention as methods for dimension reduction. Asymptotic results as well as optimality properties under certain models have been

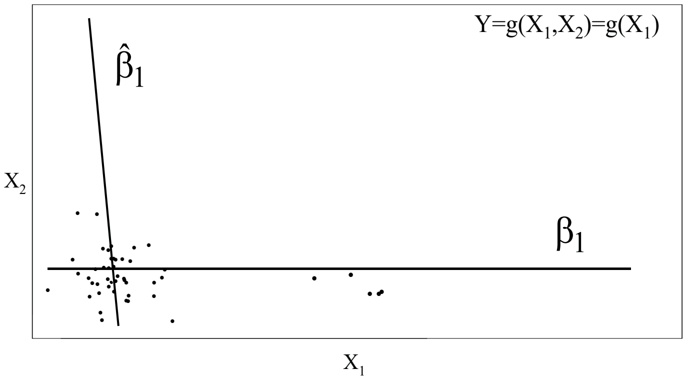


Figure 11: Estimation of dr direction by SIR for a sample containing outliers

proved.

Since SIR is meant as a first step prior to the further analysis of the data, it is very important that this method is robust against small model deviations and against the occurrence of (a few) errors in the data. Unfortunately, this is not the case, as can be seen in the following simple, two-dimensional example (see Gather et al. 2001a,b for details).

Example 4.2. *In the situation of a nonparametric regression model as described above, assume that* $Y = g(\boldsymbol{X}) = g(X_1, X_2) = f(X_1)$*. That means,* $\mathcal{B} = span(\boldsymbol{\beta}_1) = span((1,0)^T)$ *and* $K = 1$*. Consider a sample of a bivariate normally distributed regressor variable* $(X_1, X_2)^T$*, containing a few very far outlying observations, lying in the direction of* $\boldsymbol{\beta}_1$*. If we further assume* K *to be known, estimating* $\boldsymbol{\beta}_1$ *by SIR can lead to a direction* $\widehat{\boldsymbol{\beta}}_1$ *which is almost orthogonal to the true* $\boldsymbol{\beta}_1$*. Figure 11 shows an example of such a situation. If, on the other hand, we assume the more realistic situation that also* K *itself is* *unknown, for the data of Figure 11 we estimate $\widehat{K} = 0$ which means that in this case we would not even conclude that Y is related to $\boldsymbol{X}$ (compare Gather et al., 2001b).*

This phenomenon of SIR being extremely sensitive against outliers can also be shown very generally. For contaminated samples with outliers tending to lie infinitely far away from the main part of the data, the estimated dr directions tend to become orthogonal to the true dr directions in the limit (Gather et al., 2001b). In this sense, SIR can break down. One main reason for this behaviour of SIR lies in the principal component analysis based on the classical, nonrobust empirical covariance matrix. If we replace all nonrobust estimators in the SIR procedure by robust ones, meaning that we also perform a robust principal component analysis (e.g. Croux and Haesbroeck, 2000), we can improve the performance. Gather et al. (2001a) introduce such a robustified version of SIR called dimension adjustment method (DAME) and show the improved performance in samples with outliers. Figure 12 illustrates the results of a simulation study for $d = 10$, $K = 1$, $g(\boldsymbol{X}, \varepsilon) = f(X_1) + \varepsilon = X_1 + 0.1\widetilde{\varepsilon}$, $X_1, \ldots, X_{10}, \widetilde{\varepsilon}$ i.i.d. $N(0,1)$ variables, and 10% of the data contaminated by adding $2\sqrt{\chi^2_{d;0.999}}$ to the first component. For several choices of a free parameter H (the number of slices), boxplots of the absolute value of the correlation between true and estimated dr direction are shown for both procedures. The boxplots clearly demonstrate that in the outlier situation DAME is able to estimate the dr direction almost correctly (abs. correlation near one) far more often than SIR. One

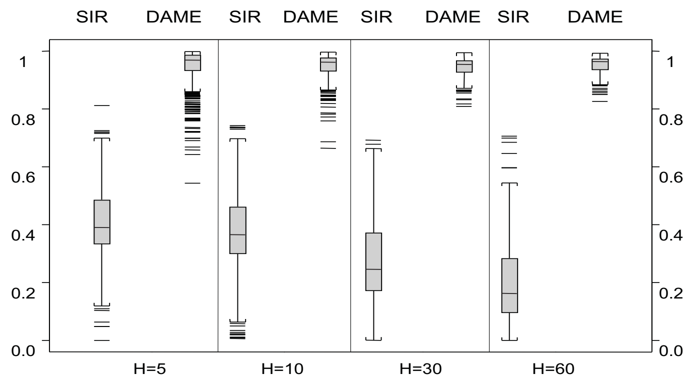


Figure 12: Simulation results: boxplots of absolute value of correlation between true and estimated dr direction by SIR and DAME for samples containing outliers

main reason for this is the use of robust estimators of the covariance matrix in the steps of DAME (for details see Gather et al., 2001a) which shows again the importance of good robust covariance estimators for high-dimensional data.

5 Conclusion

The problem of analyzing and modeling high-dimensional data is becoming more and more compelling. We are confronted with such data in many fields of modern research, in the life sciences as well as in sociology and economics, to just mention some. The reason for this can be found in the growing complexity of open problems as well as in growing possibilities of storage and computer power.

We have seen that, already when dealing with simple model structures and simple tasks for high-dimensional data, we are facing special problems. Some of them are fundamental in nature, coming with the inherent curse of dimensionality. Some are algorithmic ones due to the complexity of connected mathematical, e.g. optimization problems.

Some special solutions for partial problems are available, especially under the assumption of an underlying elliptically symmetric distribution. Often, these solutions yield new mathematical and statistical problems such as the claim for more robust methods for dimension reduction or the task of constructing "feasible" covariance matrix estimates for high-dimensional problems.

Acknowledgements

The financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of complexity in multivariate data structures") is gratefully acknowledged.

References

[1] *N. Amenta, M. Bern, D. Eppstein and S.-H. Teng:* Regression depth and center points, Discrete Comput. Geom. **23** (2000), 305–323.

[2] *D. Asimov:* The grand tour: A tool for viewing multidimensional data, SIAM J. Sci. Statist. Comput. **6** (1985), 128–143.

[3] *V. Barnett and T. Lewis:* Outliers in statistical data, 3rd ed., Wiley, New York 1994.

[4] *M. Bauer, U. Gather and M. Imhoff:* Analysis of high dimensional data from intensive care medicine (in *R. Payne and P. Green (eds.):* Proceedings in Computational Statistics, Physica, Heidelberg 1998), 185-190.

[5] *C. Becker:* The size of the largest nonidentifiable outlier as a performance criterion for multivariate outlier identification rules: The case of high-dimensional data (in *J.G. Bethlehem and P.G.M. van der Heijden (eds.):* COMPSTAT 2000. Proceedings in computational statistics, Physica, Heidelberg 2000), 211-216.

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

[7] *C. Becker and U. Gather:* The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules, to appear in Comput. Statist. and Data Anal. **36** (2001), 119–127.

[8] *R.E. Bellman:* Adaptive control processes, Princeton University Press 1961.

[9] *A. Christmann, P. Fischer and T. Joachims:* Comparison between the regression depth method and the support vector machine to approximate the minimum number of misclassifications, Technical report 53/00, SFB 475, University of Dortmund (2000).

[10] *R.D. Cook:* Principal Hessian directions revisited, J. Amer. Statist. Assoc. **93** (1998), 84–100.

[11] *R.D. Cook and S. Weisberg:* Comment on "Sliced inverse regression for dimension reduction" by K.-C. Li, J. Amer. Statist. Assoc. **86** (1991), 328–332.

[12] *C. Croux and G. Haesbroeck:* Principal component analysis based on robust estimators of the covariance or correlation matrix: Influence functions and efficiencies, Biometrika **87** (2000), 603–618.

[13] *P.L. Davies:* Asymptotic behaviour of S-estimates of multivariate location parameters and dispersion matrices, Ann. Statist. **15** (1987), 1269–1292.

[14] *P.L. Davies and U. Gather:* The identification of multiple outliers (with discussion), J. Amer. Statist. Assoc. **88** (1993), 782–801.

[15] *D.L. Donoho:* Breakdown Properties of Multivariate Location Estimators, Ph.D. Qualifying Paper, Department of Statistics, Harvard University 1982.

[16] *J.H. Friedman:* Multivariate adaptive regression splines (with discussion), Ann. Statist. **19** (1991), 1–141.

[17] *J.H. Friedman:* An overview of predictive learning and function approximation (in *V. Cherkassky, J.H. Friedman and H. Wechsler (eds.):* From statistics to neural networks. Theory and pattern recognition applications, Springer, Berlin 1994), 1–61.

[18] *J.H. Friedman:* Greedy function approximation: A gradient boosting machine, http://www-stat.stanford.edu/~jhf/ftp/trebst.ps (1999a).

[19] *J.H. Friedman:* Stochastic gradient boosting, http://www-stat.stanford.edu/~jhf/ftp/stobst.ps (1999b).

[20] *J.H. Friedman and W. Stuetzle:* Projection pursuit regression, J. Amer. Statist. Assoc. **76** (1981), 817–823.

[21] *U. Gather, T. Hilker and C. Becker:* A robustified version of sliced inverse regression, to appear in Proc. Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics, Ascona, 23.05.-28.05.1999 (2001a).

[22] *U. Gather, T. Hilker and C. Becker:* A note on outlier sensitivity of sliced inverse regression, submitted for publication (2001b).

[23] *D.J. Hand:* Data mining: Statistics and more? Amer. Statist. **52** (1998), 112–118.

[24] *M. Hausmann, C. Becker, U. Gather, O. Güntürkün:* Functional cerebral asymmetries during the menstrual cycle: A cross sectional and longitudinal analysis, submitted for publication 2001.

[25] *P.J. Huber:* Robust statistics, Wiley, New York 1981.

[26] *P.J. Huber:* Projection pursuit, Ann. Statist. **13** (1985), 435–525.

[27] *I.T. Jolliffe:* Principal component analysis, Springer, New York 1986.

[28] *D.N. Lawley and A.E. Maxwell:* Factor analysis as a statistical method, 2nd ed., Butterworths, London 1971.

[29] *G. Li and Z. Chen:* Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo, J. Amer. Statist. Assoc. **80** (1985), 759–766. Correction: J. Amer. Statist. Assoc. **80** (1985), 1084.

[30] *K.-C. Li:* Sliced inverse regression for dimension reduction (with discussion), J. Amer. Statist. Assoc. **86** (1991), 316–342.

[31] *K.-C. Li:* On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma, J. Amer. Statist. Assoc. **87** (1992), 1025–1039.

[32] *K. Pearson:* On lines and planes of closest fit to systems of points in space, Philos. Magazine **2**, Ser. 6 1901, 559–572.

[33] *G. Pison, P.J. Rousseeuw, P. Filzmoser and C. Croux:* A robust version of principal factor analysis (in *J.G. Betlehem and P.G.M. van der Heijden (eds.):* COMPSTAT 2000. Proceedings in computational statistics, Physica, Heidelberg 2000), 385–390.

[34] *D.M. Rocke:* Robustness properties of S-estimators of multivariate location and shape in high dimension, Ann. Statist. **24** (1996), 1327–1345.

[35] *P.J. Rousseeuw:* Multivariate estimation with high breakdown point (in *W. Grossmann, G. Pflug, I. Vincze and W. Wertz (eds.):* Mathematical statistics and applications, Reidel, Dordrecht 1985), 283–297.

[36] *P.J. Rousseeuw and M. Hubert:* Regression depth, J. Amer. Statist. Assoc. **94** (1999), 388–402.

[37] *P.J. Rousseeuw and A.M. Leroy:* Robust regression and outlier detection, Wiley, New York 1987.

[38] *P.J. Rousseeuw and B.C. van Zomeren:* Unmasking multivariate outliers and leverage points, J. Amer. Statist. Assoc. **85** (1990), 633–639.

[39] *W.A. Stahel:* Breakdown of covariance estimators, Research Report **31**, Fachgruppe für Statistik, ETH Zürich 1981.

[40] *V. Vapnik:* Statistical learning theory, Wiley, New York 1998.

[41] *B. Warner and M. Misra:* Understanding neural networks as statistical tools, Amer. Statist. **50** (1996), 284–293.

[42] *H. White:* Learning in neural networks: A statistical perspective, Neural Comp. **1** (1989), 425–464.

Ursula Gather, Claudia Becker

Fachbereich Statistik

Universität Dortmund

D-44221 Dortmund

gather@statistik.uni-dortmund.de

cbecker@statistik.uni-dortmund.de

Eingegangen 2001