A Note on Outlier Sensitivity of Sliced Inverse Regression

Ursula Gather, Torsten Hilker, and Claudia Becker

Work address:	Department of Statistics University of Dortmund D-44221 Dortmund, Germany
E-mail addresses:	gather@statistik.uni-dortmund.de tumhilker@aol.com cbecker@statistik.uni-dortmund.de
Corresponding Author:	Prof. Dr. Ursula Gather Department of Statistics, University of Dortmund D-44221 Dortmund, Germany
Phone:	+49 - 231 - 755 3110
Fax:	+49 - 231 - 755 5305

Running title: Outlier Sensitivity of SIR

A Note on Outlier Sensitivity of Sliced Inverse Regression

Ursula Gather, Torsten Hilker, and Claudia Becker

Abstract

Sliced Inverse Regression (SIR) is a promising technique for the purpose of dimension reduction. Several properties of this method have been examined already, but little attention has been paid to robustness aspects. In this article, we focus on the sensitivity of SIR to outliers and show in what sense and how severely SIR can be influenced by outliers in the data.

AMS 1991 Subject Classifications: Primary 62F35; secondary 62H12.

Key words and phrases: Dimension reduction; Outliers; Robustness.

1 Introduction

Many statistical methods suffer from the so-called "curse of dimensionality". As an example, consider classical nonparametric regression techniques which are used if a flexible modelling of a completely unknown functional relationship is the goal of the data analysis. It is well known that such methods (especially nearest neighbor techniques) work successfully if the explanatory variable is low-dimensional. The theoretical concepts can be transferred to the situation of high-dimensional explanatory variables, but it is almost impossible to use the resulting methods practically. The reason for this is the well-known curse of dimensionality, i.e. samples of realistic size cannot fill the high-dimensional euclidean space densely enough with observations. Consequently, locally, there are not sufficiently many observations to keep both the bias, and the variance of the classical nonparametric regression estimates small.

One possibility to overcome this problem is the construction of methods which are especially designed for the situation of high-dimensional explanatory variables such as Projection Pursuit Regression (Friedman and Stuetzle, 1981) and MARS (Friedman, 1991).

Another way is offered by techniques for reducing the dimension of the regressor space in a first step before applying then usual nonparametric regression techniques. Led by this idea, Li (1991) considers the following model for dimension reduction

$$Y = f(\boldsymbol{\beta}_1'\boldsymbol{X}, \ldots, \boldsymbol{\beta}_K'\boldsymbol{X}, \varepsilon), \tag{1.1}$$

where Y denotes a real-valued response variable, $\boldsymbol{X}$ is the d-dimensional random vector of explanatory variables, and ε is a random error independent of $\boldsymbol{X}$. The link function f is assumed to be completely unknown. This model applies as well to situations with discrete or even binary response Y, where we may have to put additional conditions upon f (see Chen and Li, 2001, Cook and Lee, 1999).

The space $\mathcal{B}$, denoting the span of the unknown directions $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$, is called effective dimension reduction (edr) space, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ are called edr directions.

We assume that $\mathcal{B}$ equals the intersection of all possible edr spaces and at the same time is an edr space itself. This means that $\mathcal{B}$ is the unique edr subspace of minimal dimension for which 1.1 holds, and hence $\mathcal{B}$ coincides with the central dimension reduction subspace in the sense of Cook (1994, 1996, 1998a,b), and K, the dimension of $\mathcal{B}$, is called the structural dimension (see also Cook and Critchley, 2000, Cook and Yin, 2001). For the aim of dimension reduction, the edr space $\mathcal{B}$ has to be estimated from the data. For this purpose Li (1991) suggests a method called Sliced Inverse Regression (SIR), which uses the fact that under some mild conditions concerning the distribution of $\boldsymbol{X}$, the centered inverse regression curve $E(\boldsymbol{X}|Y) - E(\boldsymbol{X})$ falls almost surely into a subspace spanned by $\boldsymbol{\Sigma}\boldsymbol{\beta}_1, \ldots, \boldsymbol{\Sigma}\boldsymbol{\beta}_K$, where $\boldsymbol{\Sigma}$ denotes the covariance matrix of $\boldsymbol{X}$, cf. Li (1991, Theorem 3.1). Estimating the edr directions based upon obervations $(y_i, \boldsymbol{x}_i^T)^T$, $i = 1, \ldots, n$, of $(Y, \boldsymbol{X}^T)^T$ is then carried out in five steps, yielding estimators $\widehat{\boldsymbol{\beta}}_k$, $k = 1, \ldots, K$, assuming for the moment that K is known.

The estimation scheme and the application of SIR are discussed in detail in Li (1991) and in Chen and Li (1998). SIR has been investigated in several articles with emphasis on its asymptotic properties (Li, 1991; Hsing and Carroll, 1992; Kötter, 1996; Zhu and Fang, 1996) and on the estimation of the dimension K of the edr space in a pre-step of SIR (Li, 1991; Schott, 1994; Ferré, 1997, 1998). Since design conditions concerning the distribution of $\boldsymbol{X}$ are essential for the application of SIR, cf. Li (1991, Condition 3.1), this aspect is another main topic in the discussion, cf. Cook and Weisberg (1991), Härdle and Tsybakov (1991), Hall and Li (1993), Cook and Nachtsheim (1994). Carroll and Li (1992) use SIR in a general nonlinear regression model with measurement error in the predictor variables. Aragon and Saracco (1997) consider SIR in the situation of small sample sizes, and Saracco (1999) combines SIR with canonical analysis to consider linear constraints on the slope parameters. The work of Li, Wang and Chen (1999) demonstrates that the idea of SIR also applies well to models for censored data. Bura (1997) uses a multivariate linear model for the inverse regression curve. Hsing (1999) deals with the problem of determining the slices in the second step of the SIR procedure (see Section 2) and introduces a nearest neighbor technique for this purpose. Also, the case of a multidimensional response is investigated. Two nonparametric methods for testing the hypothesis $H_0 : K = 1$ versus the alternative $H_1 : K > 1$ are proposed by Sheather and McKean (1997). SIR is implemented in the package XploRe (Härdle, Klinke, and Turlach, 1995). There are also methods for estimating the edr directions based on second moments (Cook and Weisberg, 1991) and on properties of so-called principal Hessian directions (Li, 1992, Cook, 1998a).

One aspect of SIR which has not been treated thoroughly yet is the sensitivity of this method with respect to outliers in the data though some remarks on its non-robustness have been coined. Sheather and McKean (1997, p. 271) mention that "it is well known that SIR is not robust to the effect of outliers", also referring to investigations in Cook and Weisberg (1994, pp. 124-125). Both sources deal with examples where it is the choice of the reduced dimension K which is disturbed by potential outliers in a way that K is overestimated. This is not necessarily a negative effect, because outliers can also be interpreted as additional directions of interest (Cook and Critchley, 2000). In this respect, a certain sensitivity of SIR to outliers may even be helpful in detecting such spurious observations. We supplement and complete these investigations and show in Section 3.2 that on the other hand K can also be underestimated when outliers are present. Thus, outliers can prevent the procedure from detecting even one edr direction correctly when K has to be estimated.

Moreover, the robustness problem is still present when K is known or correctly estimated in a pre-step of SIR. A complete and thorough analysis of the robustness issue is still outstanding. As we are in the situation of high-dimensional random regressor observations, one cannot just "screen out bad design points" as has been coined sometimes, because it is very difficult to identify outliers in high-dimensional data (see Rocke, 1996, Becker and Gather, 1999). A sensitivity analysis of SIR in outlier situations is therefore called for and will be presented here. It turns out in Section 3.1, that in the worst case, when K is known, a dimension reducing direction orthogonal to the true one can be estimated.

2 The SIR-procedure

As mentioned above, Sliced Inverse Regression is a technique to reduce the dimension of the regressor space to avoid the curse of dimensionality in nonparametric regression. We sketch here the essential steps of the procedure. For a detailed investigation see Li (1991), Cook (1998b). In the setting of model (1.1) and in case that

K of the reduced regressor space is known, the estimation scheme of SIR is carried out in the following five steps, based upon observations $(y_i, \boldsymbol{x}_i^T)^T$, $i = 1, \ldots, n$, of $(Y, \boldsymbol{X}^T)^T$:

1. Standardization: $\boldsymbol{z}_i = \widehat{\boldsymbol{\Sigma}}^{-1/2}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$, $i = 1, \ldots, n$, where $\widehat{\boldsymbol{\Sigma}}$ and $\overline{\boldsymbol{x}}$ denote the sample covariance matrix and the sample mean of $\boldsymbol{x}_i$, $i = 1, \ldots, n$, respectively.

2. Slicing: Divide the range of y into H sets $I_1, \ldots, I_H$, where $I_1 = \{y_{(1)} \leq \ldots \leq y_{(n_1)}\}, \ldots, I_H = \{y_{(n_{H-1}+1)} \leq \ldots \leq y_{(n)}\}$. Let n_h be the number of observations of y which fall into I_h, $h = 1, \ldots, H$, and let $\hat{p}_h = n_h/n$. Construct H slices S_h, $h = 1, \ldots, H$ for the $\boldsymbol{z}_i$ according to $\boldsymbol{z}_i \in S_h \Leftrightarrow y_i \in I_h$.

3. Sample mean-vector within each slice: $\widehat{\boldsymbol{m}}_h = \sum_{y_i \in I_h} \boldsymbol{z}_i / n_h$.

4. Principal Component Analysis (PCA): Let $\widehat{\boldsymbol{V}} = \sum_{h=1}^{H} \hat{p}_h \widehat{\boldsymbol{m}}_h \widehat{\boldsymbol{m}}_h'$ and compute the eigenvalues $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_d$ and corresponding eigenvectors $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_d$ of $\widehat{\boldsymbol{V}}$ with $\widehat{\boldsymbol{\eta}}_i' \widehat{\boldsymbol{\eta}}_j = \delta_{ij}$ where δ_{ij} denotes the Kronecker symbol.

5. Retransformation: The edr directions are estimated by $\widehat{\boldsymbol{\beta}}_k = \widehat{\boldsymbol{\Sigma}}^{-1/2} \widehat{\boldsymbol{\eta}}_k$, $k = 1, \ldots, K$.

The matrix $\boldsymbol{\Sigma}$ and $\mu = E(\boldsymbol{X})$ can be used instead of $\widehat{\boldsymbol{\Sigma}}$ and $\overline{\boldsymbol{x}}$ if they are known. Because of the partition of the data into slices and the estimation of the inverse regression curve, this estimation scheme is called Sliced Inverse Regression (SIR) (Li, 1991). The inverse regression curve is thus estimated roughly by a step function, where each slice $S_h = \{\boldsymbol{z}_i : y_i \in I_h\}$ forms a step with value $\widehat{\boldsymbol{m}}_h$, $h = 1, \ldots, H$.

The directions of highest variability within the slices then yield the estimated edr directions. Thus, SIR can be seen as a procedure in between PCA just for $\boldsymbol{X}$ and Projection Pursuit Regression for $Y = g(\boldsymbol{X})$. The slices are very often determined by choosing the number H of slices beforehand and by partitioning the sample size n into slice sizes $n_1, \ldots, n_H$ with $\sum_{h=1}^{H} n_h = n$. Let $y_{(1)} \leq \cdots \leq y_{(n)}$ be the ordered observations of Y. The n_1 smallest observations of Y are used to form the first slice I_1, the next n_2 smallest observations are used for the second slice I_2, and so on. For our considerations we will adopt this procedure. Usually, slices with equal sizes are chosen. Another possibility is to use a nearest neighbor technique (Hsing, 1999). For a discussion of choosing the number of slices see Chen and Li (1998), Cook and Weisberg (1994), Ferré (1998), or Li (1991).

3 Sensitivity of SIR to outliers

As mentioned above, some investigations up to now give only hints on the non-robustness of SIR concerning the appropriate choice of the dimension K (e.g. Cook and Weisberg, 1994, pp. 124-125, Sheather and McKean, 1997). We show here that not only the estimation of K may be affected in an undesired way by outliers but also that SIR is sensitive to outliers when the reduced dimension K is known or correctly estimated.

3.1 The case of K known

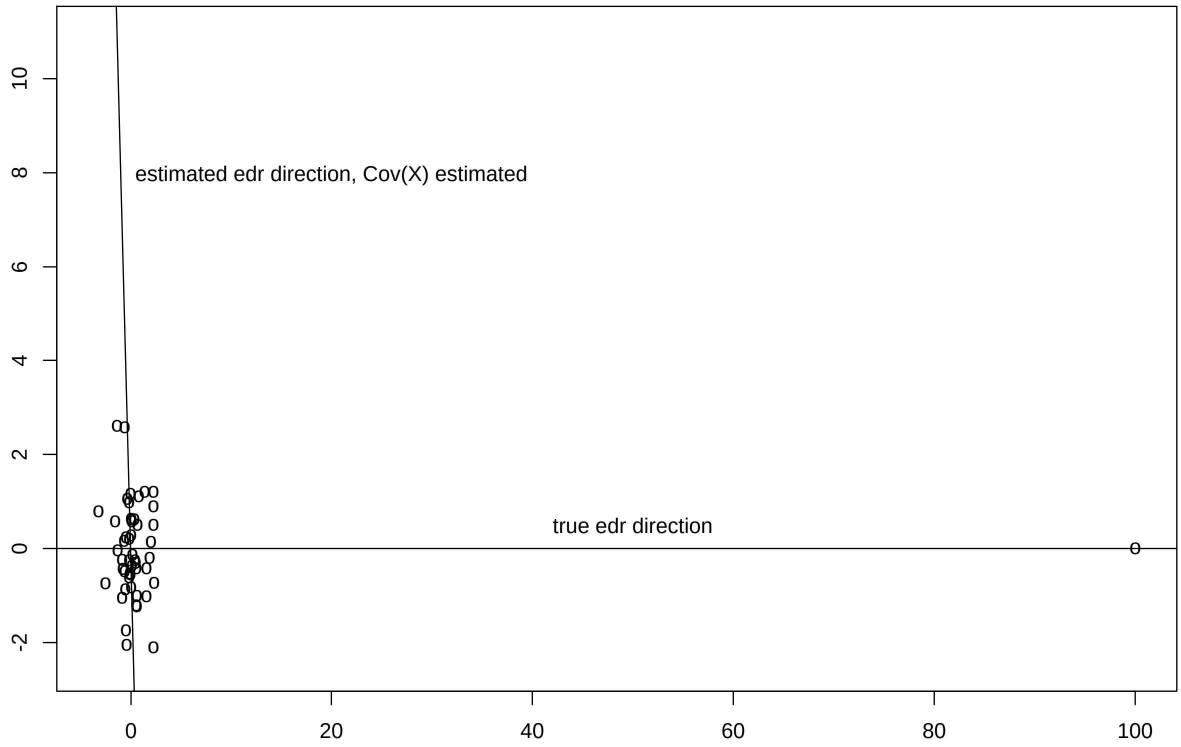
Consider the situation of model (1.1), where K is known. Take a sample of size n, and replace one observation by an arbitrarily chosen point. It can then be shown that one 'large' outlier can make it impossible to estimate all K edr directions reliably. Indeed, an edr direction can even be estimated orthogonal to the true direction as is illustrated in Figure 1. This is shown in detail now.

Remark 3.1 *The matrix* $\widehat{\boldsymbol{V}}$ *is given by* $\widehat{\boldsymbol{V}} = \widehat{\boldsymbol{\Sigma}}^{-1/2} \sum_{h=1}^{H} \widehat{p}_h (\overline{\boldsymbol{x}}_h - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_h - \overline{\boldsymbol{x}})' \widehat{\boldsymbol{\Sigma}}^{-1/2}$ *with* $\overline{\boldsymbol{x}}_h = \sum_{i:y_i \in I_h} \boldsymbol{x}_i / n_h$, $h = 1, \dots, H$. *Therefore, the estimated edr directions are also eigenvectors of*

$$\widetilde{\boldsymbol{V}} = \widehat{\boldsymbol{\Sigma}}^{-1} \sum_{h=1}^{H} \widehat{p}_h (\overline{\boldsymbol{x}}_h - \overline{\boldsymbol{x}})(\overline{\boldsymbol{x}}_h - \overline{\boldsymbol{x}})'. \tag{3.1}$$

Let $(y_1, \boldsymbol{x}_1^T)^T, \dots, (y_n, \boldsymbol{x}_n^T)^T$ be a sample of $(Y, \boldsymbol{X}^T)^T$ of size n. Without loss of generality we assume that the observations are ordered such that $y_1, \dots, y_{n_1}$ and the corresponding observations of $\boldsymbol{X}$ belong to the first slice, $y_{n_1+1}, \dots, y_{n_1+n_2}$ to the second slice and so on. Now we replace $\boldsymbol{x}_1$ by $\widetilde{\boldsymbol{x}}_{1,t} = t\boldsymbol{\beta} + \boldsymbol{u}$ with $t \in \mathrm{I\!R}$, $t > 0$, $\boldsymbol{u}, \boldsymbol{\beta} \in \mathrm{I\!R}^d$, $\boldsymbol{\beta} \neq \boldsymbol{0}$. Thus, we get a corrupted sample for each value of t and if we let t tend to infinity, the interpretation is that one observation of $\boldsymbol{X}$ is pulled away into the direction of $\boldsymbol{\beta}$. The corresponding observation y_1 of y is not altered. In general we can observe that altering observations of y only has an indirect effect on the results of SIR since those observations are only used to construct the slices (cf. Li, 1991).

Figure 1: Estimated edr direction using SIR when $Cov(\boldsymbol{X})$ is unknown: estimated direction almost orthogonal to true direction



How does the above contamination scheme change the estimation of the edr directions? The following theorem gives an answer for the case $K = 1$. A similar result holds for $K > 1$. In order to measure the performance of the estimation, we look at the correlation between $\boldsymbol{\beta}'\boldsymbol{x}$ and $\widehat{\boldsymbol{\beta}}'_{1,t}\boldsymbol{x}$, where $\boldsymbol{\beta}$ is the contaminating direction and $\widehat{\boldsymbol{\beta}}_{1,t}$ the estimated direction based on the corrupted sample for each $t > 0$.

Theorem 3.2 *Let $(y_1, \boldsymbol{x}_1^T)^T, \ldots, (y_n, \boldsymbol{x}_n)^T$ be a sample of $(Y, \boldsymbol{X}^T)^T$ of size n, and let $K = 1$. If the sample is corrupted in the above mentioned way with an arbitrarily chosen direction $\boldsymbol{\beta} \neq \boldsymbol{0}$ and if there is no $c \in \mathbb{R}$ with $\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^* = c\boldsymbol{\beta}$, then* $\lim_{t\to\infty} \widehat{\boldsymbol{\beta}}'_{1,t}\boldsymbol{\beta} = 0$.

Here, $\overline{\boldsymbol{x}}_1^* = \sum_{i=2}^{n_1} \boldsymbol{x}_i/(n_1 - 1)$ and $\overline{\boldsymbol{x}}^* = \sum_{i=2}^{n} \boldsymbol{x}_i/(n-1)$.

The Theorem tells us that under a mild condition the edr direction can be estimated almost orthogonally to an arbitrarily chosen direction $\boldsymbol{\beta}$ if there is only one bad point in the sample which is pulled away into the direction of $\boldsymbol{\beta}$. If $\boldsymbol{\beta}$ is chosen equal to $\boldsymbol{\beta}_1$, i.e. the true edr direction, we see that this direction is estimated completely wrongly. In this case, where the outlier is pulled away into the edr direction, it should not be interpreted as giving an additional direction. The proof of this Theorem is given in the Appendix.

Remark 3.3 *In the situation of Theorem 3.2, but with $K > 1$, a similar result can be shown. Let $\boldsymbol{B} = [\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K]$ be the matrix consisting of the edr directions as column vectors and let $\widehat{\boldsymbol{B}}_t = [\widehat{\boldsymbol{\beta}}_{1,t}, \dots, \widehat{\boldsymbol{\beta}}_{K,t}]$ be the matrix with the corresponding estimated directions. Then $\lim_{t\to\infty} \widehat{\boldsymbol{\beta}}'_{i,t}\boldsymbol{\beta} = 0, i = 1, \dots, K$.*
Under the additional assumptions that there exist $\varepsilon > 0$ and an unbounded and strictly increasing sequence $\{t_m\}_{m\in\mathbb{N}}$ in $\mathbb{R}$ with $\det(\widehat{\boldsymbol{B}}'_{t_m}\widehat{\boldsymbol{B}}_{t_m}) \geq \varepsilon$ for all $m \in \mathbb{N}$, and that $\boldsymbol{\beta}$ is of the form $\boldsymbol{\beta} = \boldsymbol{\Sigma}\boldsymbol{\vartheta}$, it even holds that $\lim_{m\to\infty} \varrho^2_{K,t_m} = 0$. Here, $\varrho_{1,t} \geq \cdots \geq \varrho_{K,t}$ denote the canonical correlations between $\boldsymbol{\beta}'_1\boldsymbol{x}, \dots, \boldsymbol{\beta}'_K\boldsymbol{x}$ on the one hand and $\widehat{\boldsymbol{\beta}}'_{1,t}\boldsymbol{x}, \dots, \widehat{\boldsymbol{\beta}}'_{K,t}\boldsymbol{x}$ on the other hand, where the index t expresses the dependence on the corrupted sample. (For the concept of canonical correlations see Kshirsagar (1972).) This means that the smallest canonical correlation between the true and the estimated edr directions tends to zero and therefore the estimated edr space and the true one are not the same.

A result similar to Theorem 3.2 can also be proved when $\boldsymbol{\Sigma}$ is known and has not to be estimated from the data:

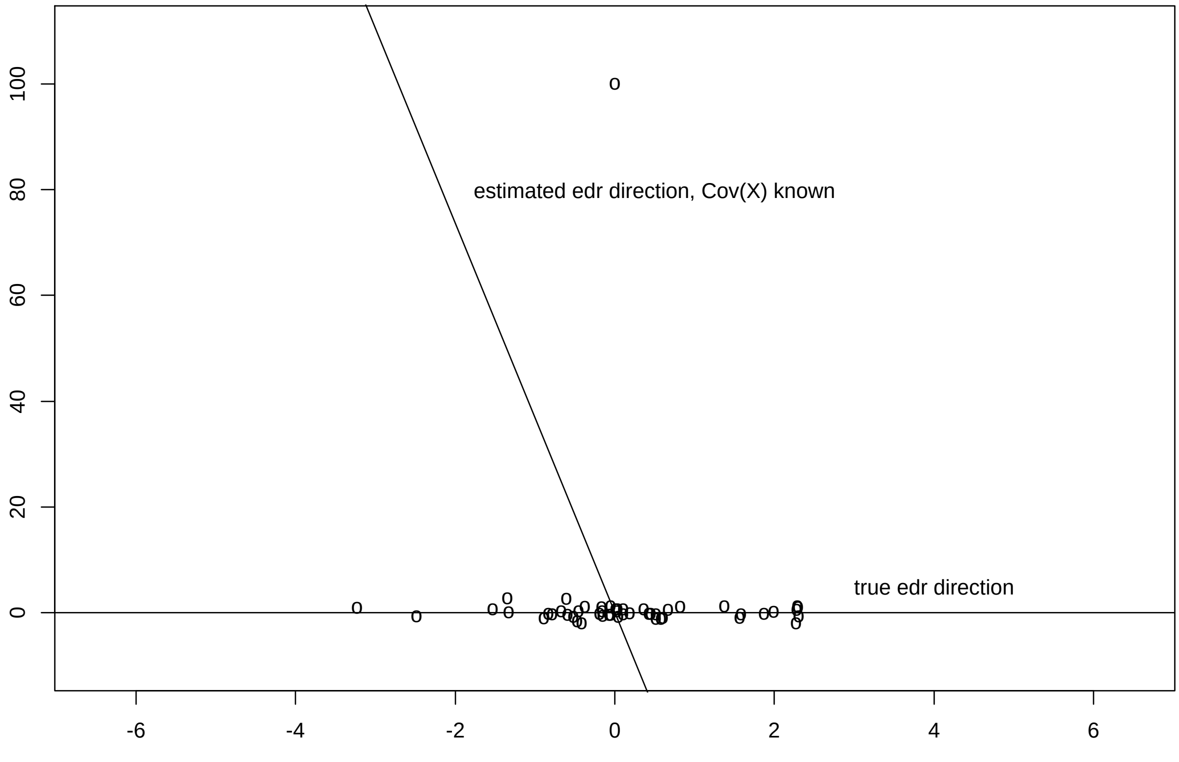
Theorem 3.4 *For known covariance matrix* $\mathbf{\Sigma}$*,*

$$\lim_{t\to\infty} \widehat{\boldsymbol{\beta}}_{1,t} = \frac{\mathbf{\Sigma}^{-1}\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}'\mathbf{\Sigma}^{-1}\boldsymbol{\beta}}} \quad \textit{and} \quad \lim_{t\to\infty} \widehat{\lambda}_{1,t} = \infty,$$

and if $\boldsymbol{\beta}'\boldsymbol{\beta}_k = 0$*,* $k = 1, \ldots, K$*, then it holds that* $\lim_{t\to\infty} \varrho^2_{K,t} = 0$.

The Theorem shows that the first estimated edr direction is extremely influenced by only one bad point which is pulled away into the direction of $\boldsymbol{\beta}$. This is illustrated in Figure 2.

Figure 2: Influence of outliers on estimated edr direction when $Cov(\boldsymbol{X})$ is known



This demonstrates again the sensitivity of SIR to one 'large' outlier. However, the remarkable difference to the case of unknown $\mathbf{\Sigma}$ lies in the choice of $\boldsymbol{\beta}$. In contrast to the situation in Theorem 3.2, $\boldsymbol{\beta}$ is orthogonal to the true edr directions in Theorem 3.4. In this case, an outlier in direction of $\boldsymbol{\beta}$ can indeed be seen as giving an additional direction which is then detected by SIR. This difference between the two cases is due to the fact that in the situation of unknown $\mathbf{\Sigma}$, the estimate of $\mathbf{\Sigma}$ (the sample covariance matrix) breaks down when one observation is pulled away to infinity. Especially, from (A.3) in the Appendix it follows that the appropriately oriented eigenvector of $\widehat{\mathbf{\Sigma}}_t$ corresponding to the largest eigenvalue tends to $\boldsymbol{\beta}$, the eigenvalue itself tends to infinity. As a consequence, $\boldsymbol{\beta}$ is an eigenvector of $\lim_{t\to\infty} \widehat{\mathbf{\Sigma}}_t^{-1}$ corresponding to the smallest possible eigenvalue 0. It follows that with increasing t all estimated edr directions corresponding to positive eigenvalues become orthogonal to $\boldsymbol{\beta}$.

3.2 Estimating K

In practical situations, we usually do not know the dimension K of the reduced regressor space. We therefore have to estimate K. For this purpose, Li (1991) proposes a method based on successive tests of $H_0^k : K = k$ vs. $H_1^k : K > k$, starting with $k = 0$. The number k for which H_0^k is not rejected for the first time then gives the estimated dimension $\widehat{K} = k$. If $\boldsymbol{X}$ is normally distributed, $n(d-K)\overline{\lambda}_{(d-K)} \sim \chi^2_{(d-K)(H-K-1)}$ asymptotically (Li, 1991, p. 321), where $\overline{\lambda}_{(d-K)}$ denotes the mean of the $(d-K)$ smallest eigenvalues of $\widehat{\boldsymbol{V}}$, $\overline{\lambda}_{(d-K)} = \sum_{i=K+1}^{d} \widehat{\lambda}_i/(d-K)$. Thus, the test statistic used in each of the aforementioned tests is $t_k := n(d-k)\overline{\lambda}_{(d-k)}$, and H_0^k is rejected if t_k exceeds an appropriate quantile of the respective χ^2 distribution with $(d-k)(H-k-1)$ degrees of freedom. Sheather and McKean (1997) report on an example where the use of this method leads to overestimating K. This might be a desired effect following Cook and Critchley (2000), who model the occurrence of possible outliers by assuming a mixture model. In their considerations, an outlier yields an additional direction, and the central subspace automatically expands to incorporate the structure generated by the outlier. In contrast to this, we consider a situation where an outlier in the $\boldsymbol{X}$ space does not necessarily give an additional direction within the regression structure. Such an outlier can lead to the effect that the dimension reduction method SIR does not only fail in finding the true edr space but that it even looses an edr direction instead of finding an extra one. We look at the performance of the estimation of K in another example.

Example 3.5 *Consider the model*

$$\boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\mathcal{I}}),\ Y = f(\boldsymbol{X}) = X_1, \tag{3.2}$$

which means that $K = 1$ and the dimension reducing direction is given as $\boldsymbol{\beta}_1 = (1, 0, \ldots, 0)^T$. The relation between Y and $\boldsymbol{X}$ is given exactly here, without introducing an error term.

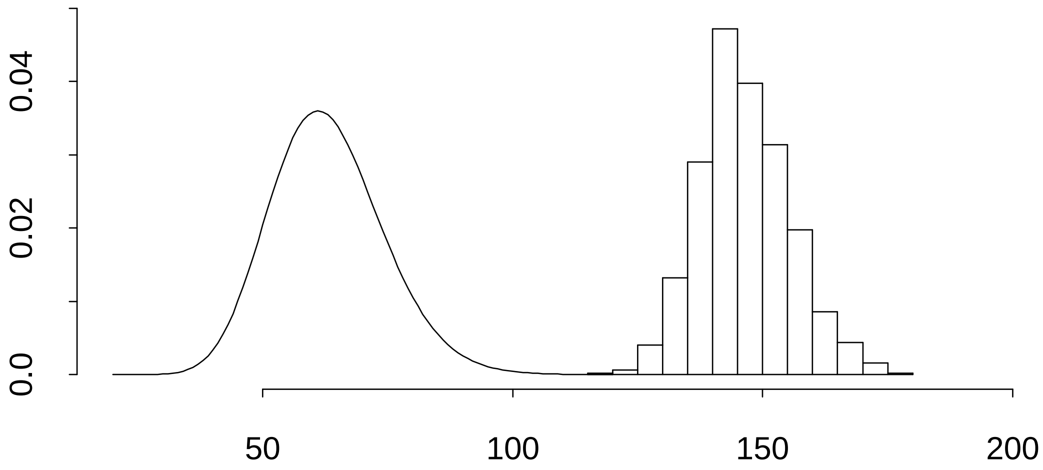
From this model, we simulate 1000 samples $(y_i, \boldsymbol{x}_i^T), i = 1, \ldots, n$ of size $n = 100$ in dimension $d = 7$. We choose $H = 10$, apply SIR and estimate K as described above. Then, carrying out each test at level $\alpha = 0.05$, we get $\widehat{K} = 1$ in 962 of 1000 cases. Thus the estimation procedure shows the desired behaviour.

Now, we take the same situation but plug in an outlier in each of the simulated samples by replacing the first coordinate of the first observation by a very extreme value, $x_{1,1} = 1,000000$. All other components of the observation $(y_1, \boldsymbol{x}_1^T)$ remain *unchanged. This means, we now have an outlier in direction of* $\boldsymbol{\beta}_1$*, the true dimension reducing direction. Then applying the same estimation method for* K *leads in only* 33 *cases to the correct* $\widehat{K} = 1$ *and to the wrong decision* $\widehat{K} = 0$ *in* 967 *of* 1000 *simulated samples.*

One can even consider a modified version of model (3.2), where $\boldsymbol{X} = (X_1, \boldsymbol{X}_2^T)^T$*, the components of* $\boldsymbol{X}$ *are independent,* $\boldsymbol{X}_2$ *is standard normal, and* X_1 *is generated stochastically by* $X_1 = X_1^* + \gamma 1,000000$*. Here,* $\gamma \in \{0,1\}$ *is assumed independent of* X_1^**,* $X_1 \sim N(0,1)$*, and* $P(\gamma = 0) = 0.999$*. Let further* $Y = X_1^*$*. We therefore have* $E(\boldsymbol{X}_2|Y) = 0$*,* $E(X_1|Y) = Y + 1000$*,* $\boldsymbol{\Sigma} = \ diag(\sigma_1, 0, \ldots, 0)$*, where* $\sigma_1 = 1 + .999(.001)(1,000000)^2$*. This leads to the conclusion that* $\boldsymbol{\Sigma}^{-1}(E(\boldsymbol{X}|Y) - E(\boldsymbol{X})) = (Y/\sigma_1, 0, \ldots, 0)^T$ *lies in the edr subspace. We see how the SIR method is influenced, although the theory behind the method still holds in this example. The very large value of* σ_1 *completely swamps the variation in* Y*.*

It can be seen from Example 3.5 that one maliciously placed outlier suffices to let the estimation method for K break down. Looking at the distribution of t_k shows what exactly happens. Figure 3 contains a histogram of the distribution of t_0 under model (3.2) with the same parameters as in Example 3.5 ($n = 100$, $d = 7$, $H = 10$). Together with this, the density of $\chi^2_{(d-k)(H-k-1)} = \chi^2_{63}$ is drawn for $k = 0$. Obviously, we have a clear separation of the distribution under the hypothesis H_0^0 and under the alternative H_1^0, leading to a good performance of the estimation procedure for K.

Figure 3: Distribution of t_0 under model (3.2): histogram of t_0 and χ^2 density



The picture changes dramatically when inserting an outlier by the same mechanism as given in Example 3.5. Again, Figure 4 shows a histogram of the distribution of t_0 together with the χ^2_{63} density, now using model (3.2) with one single outlier placed in $x_{1,1}$. In this case, it is practically impossible to distinguish between the hypothesis H_0^0 and the alternative H_1^0.

Obviously, one extreme outlier can severely influence the test statistics used in the estimation procedure for the reduced dimension K. One could therefore think of robustifying the statistics t_k to get an estimation method which is less sensitive in such situations. A first idea would be to replace the highly nonrobust mean $\overline{\lambda}_{(d-k)}$ by the very robust median $\text{med}_{(d-k)} = \text{median}(\widehat{\lambda}_{k+1}, \ldots, \widehat{\lambda}_d)$. Then $\widetilde{t}_k = n(d-k)\text{med}_{d-k}$ would be used in the respective tests of H_0^k vs. H_1^k. But Figure 5 shows that this is no solution to our problem. In the figure, a histogram of the distribution of $\widetilde{t}_0$

Figure 4: Distribution of t_0 under model (3.2) with one outlier: histogram of t_0 and χ^2 density

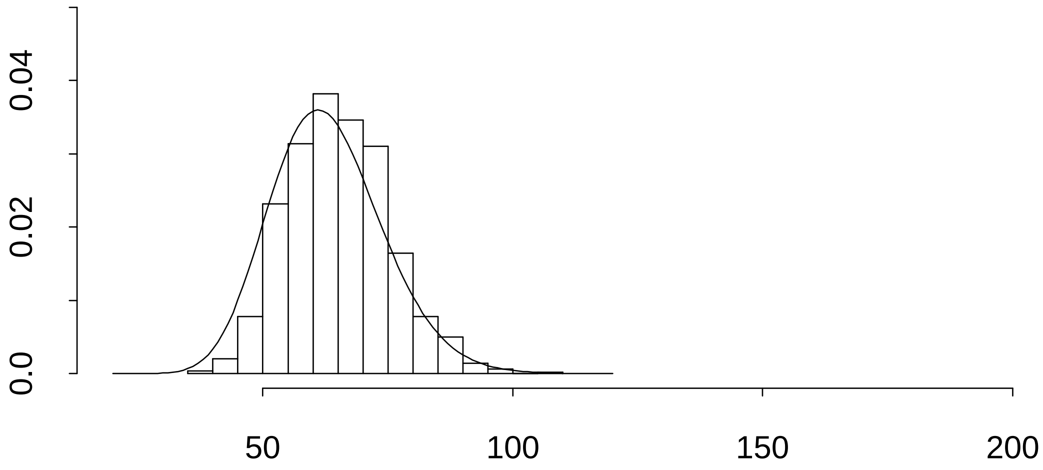
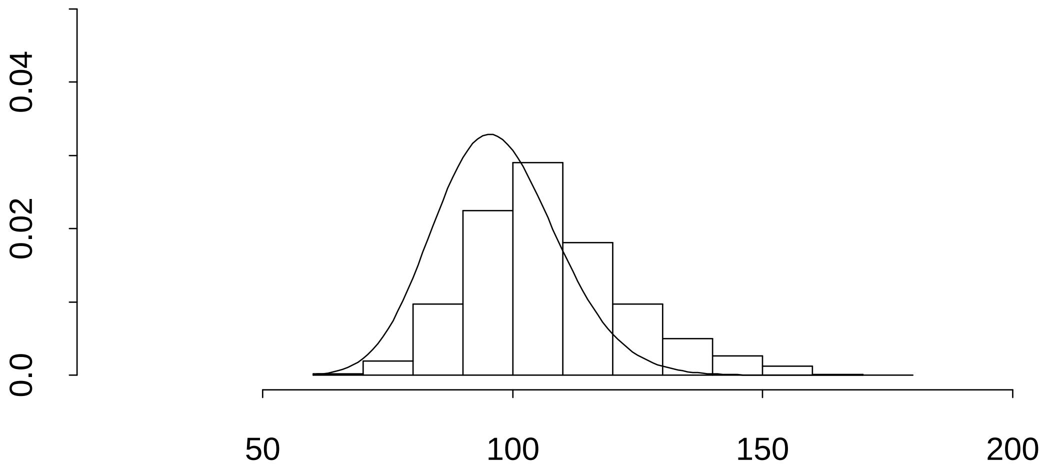


Figure 5: Density of $\widetilde{t}_0$ under H_0^0 and histogram of $\widetilde{t}_0$ under model (3.2)



under model (3.2) is printed together with the density of $\widetilde{t}_0$ under the hypothesis H_0^0. Here, the density is approximated by simulation. We see that with $\mathrm{med}_{(d-k)}$ the separation between the distribution of $\widetilde{t}_0$ under H_0^0 and under H_1^0 is not clear enough to allow a distinction between hypothesis and alternative. Thus, the use of a robustified test statistic at this stage of the method does not seem appropriate. Instead, the robustification should begin earlier, already in the five steps of the SIR procedure, which is not within the scope of this paper.

4 Conclusions

It seems obvious from the preceding section, that the sensitivity of SIR to outliers is due to the fact that nonrobust estimators are used in each step. Therefore, an immediate idea to robustify SIR is to maintain the clever estimation scheme of SIR, replacing all nonrobust estimators by alternatives with desirable robustness properties. A proposal for such a robustification of SIR goes beyond the goal of the present paper and will be given elsewhere.

Summing up it may be said that the effects of outliers on SIR cannot be neglected. The results of this method can become totally unreliable if there are a few bad outliers in the data. This is the case if the dimension K has to be estimated beforehand, but also if K is known or correctly estimated in a pre-step of SIR. In both cases, the procedure cannot cope with outliers in direction of the true edr directions. Robust estimators with suitable equivariance properties should be used to overcome this problem.

Acknowledgements

This financial support of the Deutsche Forschungsgemeinschaft (SFB 475, "Reduction of Complexity in Multivariate Data Structures") is gratefully acknowledged.

Appendix: Proofs

Proof of Theorem 3.2

We first consider two Lemmas which are needed to prove Theorem 3.2.

The sample covariance matrix of the corrupted sample is given by

$$\widehat{\boldsymbol{\Sigma}}_t = \boldsymbol{S} + \frac{n-1}{n^2} t^2 \boldsymbol{\beta}_t \boldsymbol{\beta}_t', \tag{A.3}$$

with $\boldsymbol{S} = \sum_{i=2}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}}^*)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}^*)'/n$, $\boldsymbol{\beta}_t = \boldsymbol{\beta} + (\boldsymbol{u} - \overline{\boldsymbol{x}}^*)/t$ and $\overline{\boldsymbol{x}}^* = \sum_{i=2}^n \boldsymbol{x}_i/(n-1)$.

We use the fact that $\widehat{\boldsymbol{\Sigma}}_t$ is positive definite if $\boldsymbol{S}$ is positive definite. Then we have

$$\lim_{t\to\infty} \widehat{\boldsymbol{\Sigma}}_t^{-1} = \boldsymbol{S}^{-1} - \frac{\boldsymbol{S}^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'\boldsymbol{S}^{-1}}{\boldsymbol{\beta}'\boldsymbol{S}^{-1}\boldsymbol{\beta}}, \quad \lim_{t\to\infty} t^2 \cdot \widehat{\boldsymbol{\Sigma}}_t^{-1} \boldsymbol{\beta}_t = \frac{n^2 \boldsymbol{S}^{-1}\boldsymbol{\beta}}{(n-1)\boldsymbol{\beta}'\boldsymbol{S}^{-1}\boldsymbol{\beta}}. \tag{A.4}$$

We now consider the limiting behavior of the matrix $\widetilde{\boldsymbol{V}}_t$ defined as $\widetilde{\boldsymbol{V}}$ in (3.1), but for the case of corrupted samples with t tending to infinity. Using decomposition formulae (Griffiths and Hill, 1985) we get

$$\begin{aligned}\frac{1}{t}\widetilde{\boldsymbol{V}}_t = \frac{1}{t}\boldsymbol{I}_d \quad &- \quad \frac{1}{nt}\widehat{\boldsymbol{\Sigma}}_t^{-1}\left(\boldsymbol{S}_1^* + \sum_{h=2}^H \boldsymbol{S}_h\right) \\ &- \quad \frac{1}{n}\left(1 - \frac{1}{n_1}\right)\left(t\widehat{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\beta}_t + \widehat{\boldsymbol{\Sigma}}_t^{-1}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*)\right)\left(\boldsymbol{\beta}_t + \frac{\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*}{t}\right)'\end{aligned}$$

with $\boldsymbol{S}_1^* = \sum_{i=2}^{n_1} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_1^*)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_1^*)'$, $\boldsymbol{S}_h = \sum_{i:y_i \in I_h} (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_h)(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_h)'$.

Using (A.4), yields the following Lemma:

Lemma A.1 *For positive definite* $\boldsymbol{S}$*, we have that* $\lim_{t\to\infty} \widetilde{\boldsymbol{V}}_t/t = \widetilde{\boldsymbol{V}}^*$ *with*

$$\widetilde{\boldsymbol{V}}^* = \frac{1}{n}\left(1-\frac{1}{n_1}\right)\left(\frac{\boldsymbol{\beta}'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*)}{\boldsymbol{\beta}'\boldsymbol{S}^{-1}\boldsymbol{\beta}}\boldsymbol{S}^{-1}\boldsymbol{\beta} - \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*)\right)\cdot\boldsymbol{\beta}',$$

where $\overline{\boldsymbol{x}}_1^* = \sum_{i=2}^{n_1}\boldsymbol{x}_i/(n_1-1)$ *and* $\overline{\boldsymbol{x}}^* = \sum_{i=2}^{n}\boldsymbol{x}_i/(n-1)$.

As the estimated edr directions are eigenvectors of $\widetilde{\boldsymbol{V}}_t$, we can use Lemma A.1 to find results about the limiting behavior of those directions. We need the following Lemma, the proof of which is omitted here.

Lemma A.2 *For each* $t > 0$ *let* $\boldsymbol{M}_t \in \mathrm{IR}^{d\times d}$ *be a matrix with only real eigenvalues, and* $\boldsymbol{v}_t$*,* $\|\boldsymbol{v}_t\| < \infty$*, an eigenvector of* $\boldsymbol{M}_t$ *with corresponding eigenvalue* λ_t*. If* $\lim_{t\to\infty}\boldsymbol{M}_t = \boldsymbol{u}_1\cdot\boldsymbol{u}_2'$*,* $\boldsymbol{u}_1, \boldsymbol{u}_2 \in \mathrm{IR}^d$*,* $\boldsymbol{u}_1, \boldsymbol{u}_2 \neq \boldsymbol{0}$*,* $\boldsymbol{u}_2'\boldsymbol{u}_1 = 0$ *then* $\lim_{t\to\infty}(\boldsymbol{u}_2'\boldsymbol{v}_t) = 0$.

Note that the conditions of Lemma A.2 are fulfilled for $\widetilde{\boldsymbol{V}}_t$ and $\widetilde{\boldsymbol{V}}^*$. The matrix $\widetilde{\boldsymbol{V}}^*$ possesses the structure $\widetilde{\boldsymbol{V}}^* = \boldsymbol{u}_1\boldsymbol{u}_2'$ with $\boldsymbol{u}_1'\boldsymbol{u}_2 = 0$ and $\boldsymbol{u}_2 = \boldsymbol{\beta}$ which is the direction of contamination. We now can prove Theorem 3.2.

Proof of Theorem 3.2. Without loss of generality we assume $\boldsymbol{\beta}_1'\boldsymbol{\Sigma}\boldsymbol{\beta}_1 = 1$. Note that $\widetilde{\boldsymbol{V}}^*$ possesses the structure used in Lemma A.2 with $\boldsymbol{u}_2 = \boldsymbol{\beta}$ and

$$\boldsymbol{u}_1 = \frac{1}{n}\left(1-\frac{1}{n_1}\right)\left(\frac{\boldsymbol{\beta}'\boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*)}{\boldsymbol{\beta}'\boldsymbol{S}^{-1}\boldsymbol{\beta}}\boldsymbol{S}^{-1}\boldsymbol{\beta} - \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*)\right)$$

and $\boldsymbol{u}_2'\boldsymbol{u}_1 = 0$. Moreover we have $\boldsymbol{u}_1 \neq \boldsymbol{0}$, since $\overline{\boldsymbol{x}}^* - \overline{\boldsymbol{x}}_1^*$ is not a scalar multiple of $\boldsymbol{\beta}$.

Using the fact that $\widehat{\boldsymbol{\Sigma}}_t$ is positive definite, it follows that there exist $b_0 > 0$ and $b_1 < \infty$ with $b_0 \leq \|\widehat{\boldsymbol{\beta}}_{1,t_m}\|^2 \leq b_1$.

The estimated edr direction $\widehat{\boldsymbol{\beta}}_{1,t}$ is an eigenvector of $\widetilde{\boldsymbol{V}}_t$ and consequently of $\widetilde{\boldsymbol{V}}_t/t$ for each $t > 0$. Since $\boldsymbol{u}_2 = \boldsymbol{\beta}$ we have

$$\lim_{t\to\infty} \widehat{\boldsymbol{\beta}}'_{1,t_m} \boldsymbol{\beta} = 0. \qquad \square$$

References

Y. Aragon, J. Saracco (1997) Sliced Inverse Regression (SIR): An Appraisal of Small Sample Alternatives to Slicing, *Comput. Statist.* **12**, 109–130.

C. Becker, U. Gather (1999) The Masking Breakdown Point of Multivariate Outlier Identification Rules, *J. Amer. Statist. Assoc.* **94**, 947–955.

E. Bura (1997) Dimension Reduction via Parametric Inverse Regression, *L_1-Statistical Procedures and Related Topics*, IMS Lecture Notes **31**, 215–228.

R.J. Carroll, K.-C. Li (1992) Measurement Error Regression With Unknown Link: Dimension Reduction and Data Visualization, *J. Amer. Statist. Assoc.* **87**, 1040–1050.

C.-H. Chen, K.-C. Li (1998) Can SIR be as Popular as Multiple Linear Regression? *Statist. Sinica* **8**, 289–316.

C.-H. Chen, K.-C. Li (2001) Generalization of Fisher's Linear Discriminant Analysis via the Approach of Sliced Inverse Regression, *J. Korean Statist. Soc.* **30**, 193–217.

R.D. Cook (1994) On the Interpretation of Regression Plots, *J. Amer. Statist. Assoc.* **89**, 177–189.

R.D. Cook (1996) Graphics for Regressions With a Binary Response, *J. Amer. Statist. Assoc.* **91**, 983–992.

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

R.D. Cook (1998b) *Regression Graphics: Ideas for Studying Regressions through Graphics*, Wiley, New York.

R.D. Cook, F. Critchley (2000) Identifying Regression Outliers and Mixtures Graphically, *J. Amer. Statist. Assoc.* **95**, 781–794.

R.D. Cook, H. Lee (1999) Dimension-Reduction in Binary Response Regression, *J. Amer. Statist. Assoc.* **94**, 1187–1200.

R.D. Cook, C.J. Nachtsheim (1994) Reweighting to Achieve Elliptically Contoured Covariates in Regression, *J. Amer. Statist. Assoc.* **89**, 592–599.

R.D. Cook, S. Weisberg (1991) Comment on "Sliced Inverse Regression for Dimension Reduction" by K.C. Li, *J. Amer. Statist. Assoc.* **86**, 328–332.

R.D. Cook, S. Weisberg (1994) *An Introduction to Regression Graphics*, Wiley, New York.

R.D. Cook, X. Yin (2001) Dimension Reduction and Visualization in Discriminant Analysis (with discussion), *Austral. & New Zeal. J. Statist.* **43**, 147–199.

L. Ferré (1997) Dimension Choice for Sliced Inverse Regression Based on Ranks, *Student* **2**, 95–108.

L. Ferré (1998) Determining the Dimension in Sliced Inverse Regression and Related Methods, *J. Amer. Statist. Assoc.* **93**, 132–140.

J.H. Friedman (1991) Multivariate Adaptive Regression Splines, Invited paper with discussion, *Ann. Statist.* **19**, 1–141.

J.H. Friedman, W. Stuetzle (1981) Projection Pursuit Regression, *J. Amer. Statist. Assoc.* **76**, 817–823.

P. Griffiths, I.H. Hill (eds.) (1985) *Applied Statistics Algorithms*, Royal Statistical Society, Ellis Horwood Limited.

W. Härdle, A.B. Tsybakov (1991) Comment on "Sliced Inverse Regression for Dimension Reduction" by K.C. Li, *J. Amer. Statist. Assoc.* **86**, 333-335.

W. Härdle, S. Klinke, B.A. Turlach (1995) *XploRe: An Interactive Computing Environment*, Springer, New York.

P. Hall, K.-C. Li (1993) On Almost Linearity of Low Dimensional Projections from High Dimensional Data, *Ann. Statist.* **21**, 867-889.

T. Hsing (1999) Nearest Neighbor Inverse Regression, *Ann. Statist.* **27**, 697–731.

T. Hsing, R.J. Carroll (1992) An Asymptotic Theory for Sliced Inverse Regression, *Ann. Statist.* **20**, 1040-1061.

T. Kötter (1996) An Asymptotic Result for Sliced Inverse Regression, *Comput. Statist.* **11**, 113-136.

A.M. Kshirsagar (1972) *Multivariate Analysis*, Marcel Dekker, New York.

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

K.-C. Li (1992) On Principal Hessian Directions for Data Visualization and Dimension Reduction: Another Application of Stein's Lemma, *J. Amer. Statist. Assoc.* **87**, 1025-1039.

K.-C. Li, J.-L. Wang, C.-H. Chen (1999) Dimension reduction for censored regression data, *Ann. Statist.* **27**, 1-23.

D.M. Rocke (1996) Robustness Properties of S-Estimators of Multivariate Location and Shape in High Dimension, *Ann. Statist.* **24**, 1327-1345.

J. Saracco (1999) Sliced Inverse Regression Under Linear Constraints, *Comm. Statist. – Theory and Methods* **28**, 2367–2393.

J.R. Schott (1994) Determining the Dimensionality in Sliced Inverse Regression, *J. Amer. Statist. Assoc.* **89**, 141-148.

S.J. Sheather, J.W. McKean (1997) A Comparison of Procedures based on Inverse Regression, L_1-*Statistical Procedures and Related Topics*, IMS Lecture Notes **31**, 271-278.

L.-X. Zhu, K.-T. Fang (1996) Asymptotics for Kernel Estimate of Sliced Inverse Regression, *Ann. Statist.* **24**, 1053-1068.