IMS Collections
Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jurečková
Vol. 7 (2010) 75–83

DOI: 10.1214/10-IMSCOLL708

MCD-RoSIS – A robust procedure for variable selection*

Charlotte Guddat[1] and Ursula Gather[1] and Sonja Kuhnt[1]

TU Dortmund University

Abstract: Consider the task of estimating a regression function for describing the relationship between a response and a vector of p predictors. Often only a small subset of all given candidate predictors actually effects the response, while the rest might inhibit the analysis. Procedures for variable selection aim to identify the *true* predictors. A method for variable selection when the dimension p of the regressor space is much larger than the sample size n is Sure Independence Screening (SIS). The number of predictors is to be reduced to a value less than the number of observations *before* conducting the regression analysis. As SIS is based on nonrobust estimators, outliers in the data might lead to the elimination of true predictors. Hence, a *robustified* version of SIS called RoSIS was proposed which is based on robust estimators. Here, we give a modification of RoSIS by using the MCD estimator in the new algorithm. The new procedure MCD-RoSIS leads to better results, especially under collinearity. In a simulation study we compare the performance of SIS, RoSIS and MCD-RoSIS w.r.t. their robustness against different types of data contamination as well as different degrees of collinearity.

1. Introduction

In the analysis of high dimensional data the curse of dimensionality Bellmann [1] is a phenomenon which hinders an accurate modeling of the relation between a response variable $Y \in \mathbb{R}$ and a p-dimensional vector of predictors $\boldsymbol{X} = (X_1, \ldots, X_p)^T \in \mathbb{R}^p$. There are essentially two ways to handle the problem: we either use a regression method that is able to cope with high dimensional data, or we apply a dimension reduction technique that projects the p-dimensional predictor onto a subspace of lower dimension $K \ll p$ followed by a *usual* regression procedure.

For the later approach, Li [10] proposed the model

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

where $f : \mathbb{R}^K \to \mathbb{R}$ is an unknown link function to be estimated from observations $(\boldsymbol{x}_i^T, y_i)^T$, $i = 1, \ldots, n$, and ε is an error term that is independent from $\boldsymbol{X}$. The vectors $\boldsymbol{b}_i$, $i = 1, \ldots, K$, are called effective dimension reduction (edr) directions which span a K-dimensional subspace $\mathcal{S}_{Y|\mathbf{X}}$ assumed to be the central subspace in

*This work was partially supported by the German Science Foundation (DFG, SFB 475, "Reduction of complexity in multivariate data structures", and SFB 823, "Statistical modelling of nonlinear dynamic processes").

[1]Faculty of Statistics, TU Dortmund University, 44221 Dortmund, Germany,
e-mails: guddat,gather,kuhnt@statistik.tu-dortmund.de

AMS 2000 subject classifications: Primary 62G35, 62J99.
Keywords and phrases: Variable selections, dimension reduction, regression, outliers, robust estimation.

the sense of Cook [2, 3]. Under model (1.1) the projection of $\boldsymbol{X}$ onto $\mathcal{S}_{Y|\mathbf{X}}$ captures all relevant information that is given by the original data. In this paper we further restrict the link function by assuming a linear model $Y = \boldsymbol{b}^T \boldsymbol{X} + \varepsilon$ with $\boldsymbol{b} \in \mathbb{R}^p$.

Commonly, variable selection is conducted simultaneously to the regression analysis — it is part of the *model selection* Li et al. and Cox and Snell [11, 4]. Here, we focus on variable selection as a *prestep* to the regression and assume model (1.1). A special case of dimension reduction arises if all edr directions are projections onto one component of $\boldsymbol{X}$ each. Hence, out of the p predictors at hand only K_{VS} canonical unit vectors $\boldsymbol{b}_i \in \mathbb{R}^p$, $i = 1, \ldots, K_{VS}$, $K_{VS} \ll p$, are classified as being relevant and are solely used in the following regression analysis.

These days, we face a more difficult situation than the one described above more and more often: The sample size n can be much smaller than the dimension p of the regressor space. The accomplishment of this challenge is an important part of current research. Fan and Lv [6] provide a procedure for variable selection especially for this situation. They can even show that their method *Sure Independence Screening* (SIS) possesses the *sure screening property*. That is, after the selection of $n - 1$ or $n/\log(n)$ variables by SIS, all true predictors are in the chosen subset with a very high probability when some conditions are fulfilled.

However, SIS is based on nonrobust estimators such that outliers in the data might influence the selection of predictors negatively, i. e. variables with an effect on Y are not extracted or noise variables are selected as being relevant. Hence, Gather and Guddat (2008) provide a robust version of SIS called *RoSIS — Robust Sure Independence Screening*. Here, we suggest a further modification which results in the new procedure *MCD-RoSIS* being in many situations even more robust than RoSIS and also working better under collinearity. We show this by a simulation study where we replace observations by outliers in the response as well as in the predictors and vary the sample size and the dimension of the regressor space. Also, we investigate different degrees of collinearity.

2. SIS and RoSIS

Sure Independence Screening (SIS; Fan and Lv [6]) is a procedure for variable selection that is constructed for situations with $p \gg n$. Assuming the linear model, the method is based on the determination of the pairwise covariances of each standardized predictor Z_j, $j = 1, \ldots, p$, with the response. Aim is to reduce the number of predictors to a value K_{SIS} which is smaller than the sample size n. Therefore, those variables whose pairwise covariance with Y belong to the absolutely largest, are selected for the following regression analysis.

The empirical version of $Z_j = (X_j - \mu_j)/\sigma_j$ results from the substitution of the expectation μ_j and the variance σ_j^2 of X_j by the corresponding arithmetic mean $\overline{X}_j$ and the empirical variance s_j^2, $j = 1, \ldots, p$, respectively. For the estimation of the covariance $\mathrm{Cov}(Z_j, Y)$, $j = 1, \ldots, p$, the empirical covariance is used. All these estimators are sensitive against outliers as we know. Hence, it is possible that outliers lead to an underestimation of the relation between a true predictor and Y or to an overestimation of the relation between a noise variable and Y, respectively. In the case of a strong deviation between true and estimated covariance, the elimination of a true predictor results. To avoid this, Gather and Guddat [7] introduce a robust version of SIS which is based on a robust standardization of the predictors and a robust estimation of the covariances using the Gnanadesikan–Kettenring estimator Gnanadesi and Kankettenring [8] employing the robust tau-estimate for

estimating the univariate scale Maronna and Zamar [12]. First comparisons of this new method *Robust Sure Independence Screening* (RoSIS) with SIS have shown promising results Gather and Guddat [7].

However, as previous results indicate that the Gnanadesikan-Kettenring estimator is not the best choice under collinearity for example, we suggest a version of RoSIS which employs the Minimum Covariance Determinant (MCD) estimator Rousseeuw [14] coping with this situation much better. We call this version *MCD-RoSIS* and refer to RoSIS in the following as *GK-RoSIS* for a better distinction. After a robust standardization and the estimation of the pairwise covariances by the MCD estimator the resulting values are ordered by their absolute size. Those predictors belonging to the K_{SIS} largest results are selected for the following analysis. The number K_{SIS} is to be chosen smaller than the sample size, e. g. Fan and Lv [6] suggest $K_{SIS} = n - 1$ or $K_{SIS} = n/log(n)$.

Definition 2.1. *Let* $\{(\boldsymbol{X}_1^T, Y_1)^T \ldots, (\boldsymbol{X}_n^T, Y_n^T)^T\}$ *be a sample of size* n *in* $\mathbb{R}^{p+1}$, *where* $p >> n$, *and* $K_{SIS} \in \{1, \ldots, n\}$ *given.* ***MCD-RoSIS*** *selects the variables as follows:*

(i) Robust standardization of the observations of the predictors by Median *and* MAD.
(ii) Robust estimation of the pairwise covariances $\mathrm{Cov}(Z_j, Y)$ *by* $\hat{\omega}_{rob,j} = C_{MCD}(\{z_{1,j}, \ldots, z_{n,j}\}, \{y_1, \ldots, y_n\})$, $j = 1, \ldots, p$, *by means of the* MCD estimator.
(iii) Ordering of the estimated values by their absolute size: $|\hat{\omega}_{rob,j_1}|_{(1)} \leq |\hat{\omega}_{rob,j_2}|_{(2)} \leq \ldots \leq |\hat{\omega}_{rob,j_p}|_{(p)}$.
(iv) Selection of K_{SIS} *variables:*

$$\mathcal{U} = \left\{ Z_j : |\omega_{rob,j_{K_S}}|_{(K_S)} \leq |\omega_{rob,j}|,\ 1 \leq j \leq p \right\}.$$

In the following section we examine to which extent SIS, GK-RoSIS and MCD-RoSIS are robust against large aberrant data points by means of a simulation study and compare the performance of both methods in different situations regarding the dimension p, the sample size n, the types of outliers as well as the degree of collinearity.

3. Comparison of SIS and MCD-RoSIS

In order to examine the effect of outliers on the correct selection of predictors, we simulate different *outlier scenarios*. We look at the effect of outliers in predictor variables and in the response variable while we vary the dimension p, the sample size n as well as the degree of collinearity. The following subsection contains a detailed description of the data generating processes. All simulations are carried out using the free software R (2008).

We look at three different models. The setup is the same as Fan and Lv [6] chose for checking the performance of SIS. The n observations of the p predictors $X_1, \ldots, X_p$ are generated from a multivariate normal distribution $\mathcal{N}(0, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = (\sigma_{ij}) \in \mathbb{R}^{p \times p}$ having the entries $\sigma_{ii} = 1$, $i = 1, \ldots, p$, and $\sigma_{ij} = \rho$, $i \neq j$. The observations of ε are drawn from an independent standard normal distribution. The response is assigned according to the model $Y = f(\boldsymbol{X}) + \varepsilon$ where $f(\boldsymbol{X})$ is the link function chosen as presented in Model 1 through Model 3.

$$
\begin{aligned}
&\text{Model 1:} && Y = 5X_1 + 5X_2 + 5X_3 + \varepsilon,\\
&\text{Model 2:} && Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + \varepsilon,\\
& && \text{where } \mathrm{Cov}(X_4, X_j) = \rho^{1/2}\ ,\ j = 1, 2, 3, 5, \ldots, p\\
&\text{Model 3:} && Y = 5X_1 + 5X_2 + 5X_3 - 15\rho^{1/2}X_4 + X_5 + \varepsilon\\
& && \text{where } \mathrm{Cov}(X_4, X_j) = \rho^{1/2}\ ,\ j = 1, 2, 3, 5, \ldots, p,\\
& && \text{and } \mathrm{Cov}(X_5, X_j) = 0\ ,\ j = 1, 2, 3, 4, 6 \ldots, p.
\end{aligned}
$$

The models are taken over from Fan and Lv [6] simulations. The link function in Model 1 is linear in three predictors and a noise term. The second link function includes a fourth predictor which has correlation $\rho^{1/2}$ with all the other $p-1$ candidate predictors, but is uncorrelated with the response. Hence, SIS can pick all true predictors only by chance. In the third model a fifth variable is added that is uncorrelated with the other $p-1$ predictors and that has the same correlation with Y as the noise has. Depending on ρ, X_5 has weaker marginal correlation with Y than $X_6, \ldots, X_p$ and hence has a lower priority of being selected by SIS.

We consider a dimension of $p = 100$ and 1000; the sample size is set to be $n = 50$ and 70; collinearity is varied by $\rho = 0, 0.1, 0.5, 0.9$. The number of repetitions is 200. We apply SIS, GK-RoSIS and MCD-RoSIS to each generated data set for the selection of $n-1$ variables.

For contaminating the data we replace 10% of the simulated observations by values, which are on the boundary of specific tail regions according to the notion of α-outliers Davies and Gather ([5]). For a contamination of the response we replace y_i by $f(\boldsymbol{x}) + z_{1-\alpha/2}$ with $z_{1-\alpha/2}$ the $(1-\alpha/2)$-quantile of the error distribution and $\alpha = 1 - 0.999^{\frac{1}{n}}$ depending on the sample size n, keeping x_i as it is. Concerning contamination of $\boldsymbol{X}$ we distinguish between two different directions. We place outliers in X_1- or in $X_1 + X_2 + X_3$-direction by choosing a contamination such that $\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x} = \chi^2_{0.999^{\frac{1}{n}}, p}$, with $\chi^2_{0.999^{\frac{1}{n}}, p}$ the quantiles of the χ^2-distribution with p degrees of freedom. For the X_1-direction we keep the values $\boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,p}$ and use the largest solution of the equation with respect to the first entry of $\boldsymbol{x}$ as replacement for $\boldsymbol{x}_{i,1}$. For the $X_1 + X_2 + X_3$-direction we insert $\boldsymbol{x}_{i,4}, ..., \boldsymbol{x}_{i,p}$, set the first three entries of $\boldsymbol{x}$ equal and take the largest solution as replacement for $\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, \boldsymbol{x}_{i,3}$.

As the goal of a method for variable selection is to detect the predictors which have an influence on the response a natural measure of performance is the number of correctly selected as well as the number of falsely selected predictors. As we fix the number of variables to be selected as $K_{SIS} = n - 1$ it is sufficient to look at the number of correctly selected variables.

In the following we shortly summarize the resulting performance of SIS, GK-RoSIS and MCD-RoSIS. Generally, we found that the new method MCD-RoSIS identifies all true predictors in almost 100% of the cases for all settings when the data are contaminated in one of the $\boldsymbol{X}$-directions while the classical procedure SIS fails here very often. Especially, under high collinearity or when the dimension p is large the performance of SIS is very bad. In these situations partly none of the predictors can be identified by SIS in many cases. GK-RoSIS works mostly better than SIS, but not as good as MCD-RoSIS.

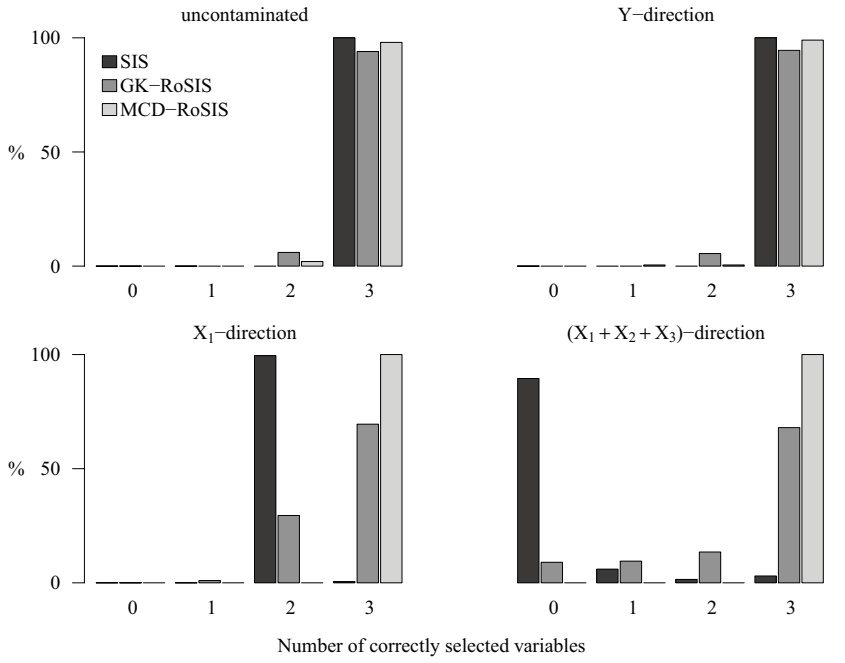

FIG 1. *SIS, GK-RoSIS and MCD-RoSIS in Model 1 with* $p = 100, n = 70, \rho = 0.5$

Comparing both procedures when the data are uncontaminated or contaminated in Y-direction we have to distinguish between the models. While for Model 1 MCD-RoSIS is only almost as good as SIS, it is generally speaking the better choice for Model 2 and 3. GK-RoSIS is rather on the same level as SIS but suffers strongly from high collinearity.

Figure 1 shows the performance of SIS GK- and MCD-RoSIS for Model 1 with parameters $p = 100$, $n = 70$ and $\rho = 0.5$. As described before, all three procedures perform similarly good for uncontaminated data and when outliers are given in the response. For the situations with outliers in $\boldsymbol{X}$ the superiority of MCD-RoSIS is obvious.

Concerning Model 2 Figure 2 shows the case of parameters $p = 100$, $n = 70$ and $\rho = 0.9$. In all data situation SIS and GK-RoSIS correctly select all predictors in around $50 - 60\%$ of the cases, whereas MCD-RoSIS has a rate of more than 95%.

In Figure 3 we find the results for Model 3 with parameters $p = 1000$, $n = 50$ and $\rho = 0.1$. This model includes a predictor that has only a very small correlation with the response. That is why SIS is not able to identify this variable X_5 even when the data are generated from the assumed model. Clearly, MCD-RoSIS finds more true predictors.

To complement the treated parameter situations, Table compares all methods, data situations and models for parameters $p = 1000$, $n = 50$ and $\rho = 0$. For all other simulations results see Guddat et al. [9].

We have seen that the MCD-RoSIS and GK-RoSIS are the better procedures for variable selection when outliers in $\boldsymbol{X}$ are present while MCD-RoSIS is at least a little weaker in the uncontaminated situations. It has also turned out that GK-

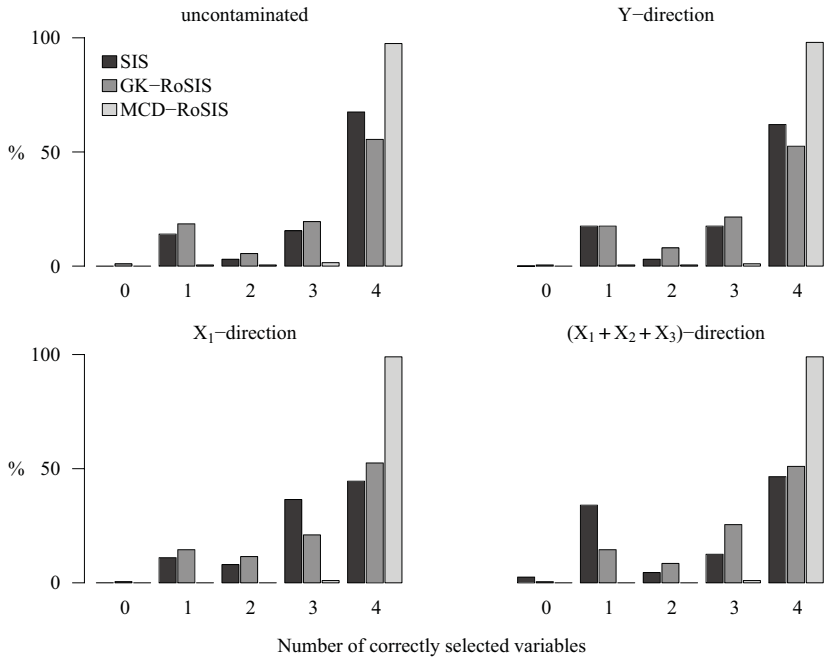

FIG 2. *SIS, GK-RoSIS and MCD-RoSIS in Model 2 with* $p = 100, n = 70, \rho = 0.9$

RoSIS suffers from collinearity as it shows inferior results in the respective situations of contamination. The reason presumably lies in the fact that the Gnandesikan–Kettenring estimator is based on univariate scale estimators. We have also observed that MCD-RoSIS is more suitable even for uncontaminated data when true predictors have only a small or no correlation with the response.

At first sight it is a little bit unexpected that the robustified procedures do not perform generally better when there is a contamination in Y-direction. The reason is that the size of α-outliers is dependent on the dimension. As the response is one dimensional, the magnitude of outlying observations in this direction is comparatively small. Hence, the application of robust estimators in the algorithm for variable selection is not beneficial yet. But the superiority of MCD-RoSIS increases along with the magnitude of the outliers. Altogether, we can conclude that MCD-RoSIS is a very good alternative for the variable selection in high dimensional settings.

4. Summary

We provide a robustified version of Sure Independence Screening (SIS) introduced by Fan and Lv [6] which is a procedure for variable selection when the number of predictors is much larger than the sample size. Aim is the reduction of the dimension to a value which is smaller than the sample size such that *usual* regression methods are applicable. We modify the algorithm by using robust estimators. To be precise, we employ Median and MAD for standardization as well as the MCD covariance estimator for the identification of the important variables. This leads to the new procedure MCD Robust Sure Independence Screening (MCD-RoSIS).

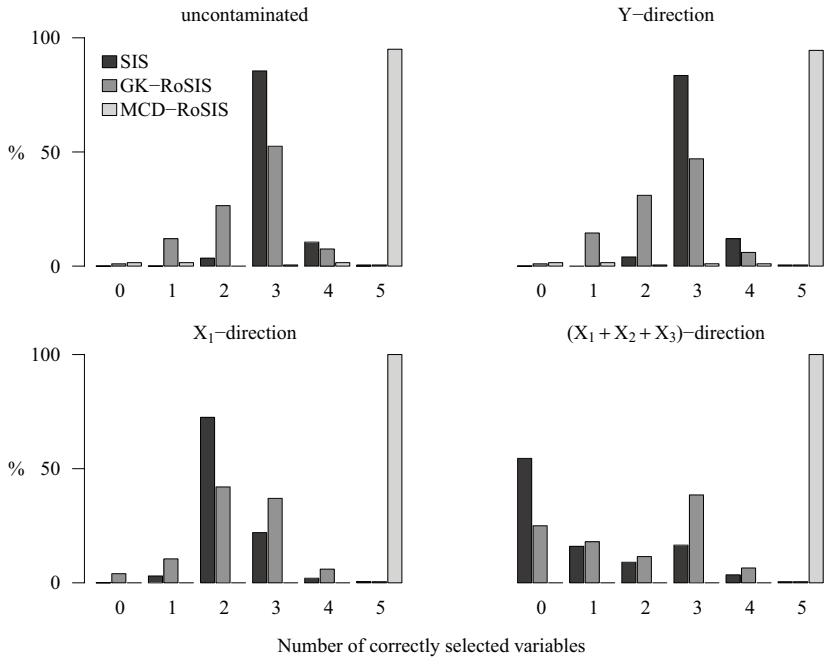

FIG 3. *SIS, GK-RoSIS and MCD-RoSIS in Model 3 with* $p = 1000, n = 50, \rho = 0.1$

In a simulation study we compare the performance of the classical procedure SIS and of the robustified versions GK- and MCD-RoSIS in different scenarios. We observe that MCD-RoSIS is the better choice for variable selection under strong contamination of the data. But we can also detect that MCD-RoSIS is at least almost as good as the classical procedure in the uncontaminated situations. GK-RoSIS is in many contaminated situations better than SIS, but it is also very sensible against collinearity. In case of predictors that have only small correlation with the response MCD-RoSIS always finds more often all true predictors even when the data are uncontaminated. Under comparatively small deviations the robustified procedure is not always the better choice. In these situations the behavior corresponds to that in the uncontaminated case. Obviously, as in other data situations the outliers must be of some size such that the use of robust estimators is profitable.

TABLE. *Simulation results for* $p = 1000,\ n = 50,\ \rho = 0$

Model 1	method	No. of correctly sel. predictors			
		0	1	2	3
uncontaminated	SIS	0.000	0.000	0.010	0.990
	GK-RoSIS	0.020	0.130	0.225	0.625
	MCD-RoSIS	0.025	0.020	0.005	0.950
Y-direction	SIS	0.000	0.000	0.015	0.985
	GK-RoSIS	0.010	0.115	0.280	0.595
	MCD-RoSIS	0.030	0.030	0.005	0.935
X_1-direction	SIS	0.000	0.005	0.865	0.130
	GK-RoSIS	0.035	0.130	0.365	0.470
	MCD-RoSIS	0.000	0.000	0.000	1.000
$(X_1 + X_2 + X_3)$-direction	SIS	0.590	0.120	0.080	0.210
	GK-RoSIS	0.245	0.175	0.115	0.465
	MCD-RoSIS	0.000	0.000	0.000	1.000

Model 2	method	No. of correctly sel. predictors				
		0	1	2	3	4
uncontaminated	SIS	0.000	0.000	0.010	0.940	0.050
	GK-RoSIS	0.015	0.130	0.230	0.605	0.020
	MCD-RoSIS	0.010	0.035	0.000	0.005	0.950
Y-direction	SIS	0.000	0.000	0.015	0.940	0.045
	GK-RoSIS	0.005	0.120	0.265	0.565	0.045
	MCD-RoSIS	0.020	0.030	0.010	0.010	0.930
X_1-direction	SIS	0.000	0.005	0.820	0.170	0.005
	GK-RoSIS	0.035	0.120	0.375	0.450	0.020
	MCD-RoSIS	0.000	0.000	0.000	0.000	1.000
$(X_1 + X_2 + X_3)$-direction	SIS	0.560	0.145	0.085	0.195	0.015
	GK-RoSIS	0.245	0.175	0.115	0.435	0.030
	MCD-RoSIS	0.000	0.000	0.000	0.000	1.000

Model 3	method	No. of correctly sel. predictors					
		0	1	2	3	4	5
uncontaminated	SIS	0.000	0.000	0.015	0.830	0.150	0.005
	GK-RoSIS	0.015	0.100	0.260	0.545	0.080	0.000
	MCD-RoSIS	0.020	0.010	0.005	0.000	0.010	0.955
Y-direction	SIS	0.000	0.000	0.025	0.825	0.145	0.005
	GK-RoSIS	0.005	0.115	0.295	0.500	0.085	0.000
	MCD-RoSIS	0.005	0.010	0.005	0.005	0.020	0.955
X_1-direction	SIS	0.000	0.010	0.735	0.215	0.040	0.000
	GK-RoSIS	0.050	0.075	0.430	0.385	0.060	0.000
	MCD-RoSIS	0.000	0.000	0.000	0.000	0.000	1.000
$(X_1 + X_2 + X_3)$-direction	SIS	0.495	0.200	0.080	0.175	0.050	0.000
	GK-RoSIS	0.205	0.190	0.160	0.380	0.065	0.000
	MCD-RoSIS	0.000	0.000	0.000	0.000	0.000	1.000

References

[1] BELLMAN, R. E. (1961). *Adaptive Control Processes.* Princeton University Press.

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

[3] COOK, R. D. (1998). Regression Graphics: Ideas for Studying Regressions Through Graphics. Wiley, New York.

[4] COX, D. R., SNELL, E. J. (1974). The Choice of Variables in Observational Studies. *Appl. Statist.* **23** 51–59.

[5] DAVIES, P. L. AND GATHER, U. (1993). The Identification of Multiple Outliers (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **88** 782–792.

[6] FAN, J. Q. AND LV, J. (2008). Sure Independence Screening for Ultrahigh Dimensional Feature Space (with discussion and rejoinder). *J. Roy. Stat. Soc. B* **70** 849–911.

[7] Gather, U. and Guddat, C. (2008). Comment on "Sure Independence Screening for Ultrahigh Dimensional Feature Space" by Fan, J.Q. and Lv, J. *J. Roy. Stat. Soc. B* **70** 893–895.

[8] Gnanadesikan, R., Kettenring, J. (1972). Robust Estimates, Residuals, and Outlier Detection With Multiresponse Data. *Biometrics* **28** 81–124.

[9] Guddat, C., Gather, U., and Kuhnt, S. (2010). MCD-RoSIS - A Robust Procedure for Variable Selection. *Discussion Paper, SFB 823, TU Dortmund, Germany.*

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

[11] Li, L., Cook, R. D. and Nachtsheim, C. J. (2005). Model-free Variable Selection. *J. Roy. Stat. Soc. B* **67** 285–299.

[12] Maronna, R. A., Zamar, R. H. (2002). Robust Estimates of Location and Dispersion for High-dimensional Datasets. *J. Amer. Statist. Assoc.* **44** 307–317.

[13] R Development Core Team (2008). R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria, ISBN 3-900051-07-0, URL http://www.R-project.org.

[14] Rousseeuw, P. J. (1984). Least Median of Squares Regression. *J. Amer. Statist. Assoc.* **84** 871–880.