Statistics and Its Interface Volume 6 (2013) 45–52

A note on robust kernel inverse regression

Yuexiao Dong*, Zhou Yu and Yizhi Sun

As a useful tool for sufficient dimension reduction, kernel inverse regression (KIR) can effectively relieve the curse of dimensionality by finding linear combinations of the predictor that contain all the relevant information for regression. However, KIR is sensitive to outliers, and will fail when the predictor distribution is heavy-tailed. In this paper, we discuss robust variations of KIR that do not have such limitations. The effectiveness of our proposed methods is demonstrated via simulation studies and an application to the automobile price data.

AMS 2000 subject classifications: Primary 60K35, 60K35; secondary 60K35.

Keywords and phrases: Elliptical distribution, Kernel inverse regression, Permutation test, Sufficient dimension reduction.

1. INTRODUCTION

High-dimensional data are becoming more and more prevalent nowadays due to the development of science and technology. How to reduce the dimensionality of the data while keeping the relevant information poses challenges for statisticians. Let X be a p-dimensional predictor and Y be a 1-dimensional response. [16] considered $Y = g(\beta^T X, \varepsilon)$, where $g(\cdot)$ is an unknown link function, $\beta \in \mathbb{R}^{p\times d}$ contains information about the relevant predictors, and ε is the random error independent of X. Many useful semiparametric models fall within this framework, such as single/multi-index models, logistic regression, general additive models, Cox's proportional hazard model, etc. An important feature of this model is that Y is independent of X conditioning on $\beta^T X$. To find β with the smallest column space such that $Y \perp\!\!\!\perp X|\beta^T X$, [1] introduced the notion of sufficient dimension reduction (SDR). This smallest column space is called the central space, and denoted as $\mathcal{S}_{Y|X}$. The dimension of $\mathcal{S}_{Y|X}$ is called the structural dimension.

Without loss of generality, assume $E(X) = 0$ and $\text{Var}(X) = I_p$. Denote β as the basis of $\mathcal{S}_{Y|X}$. In the seminar paper of sliced inverse regression (SIR; [16]), it was shown that the conditional mean of the inverse regression belongs to the central space, or $E(X|Y) \in \mathcal{S}_{Y|X}$. Thus we can effectively reduce the predictor dimensionality without knowing the form of the link function $g(\cdot)$. Instead, the so-called linear conditional mean (LCM) assumption is imposed on thepredictor distribution, which requires $E(X|\beta^T X)$ to be a linear function of $\beta^T X$. When X is elliptically-distributed, the LCM assumption is satisfied [8]. Without knowing true β in practice, [3] suggested transformation or reweighting of the predictor such that the predictor becomes approximately elliptical. For non-elliptically distributed predictor X, please refer to [14, 7].

We will focus on elliptically-distributed predictor in this paper. The density function of an elliptically contoured predictor $X \in \mathbb{R}^p$ has the form of

$$f(X) = |\Gamma|^{-1/2}\ell\left(\|X-\mu\|^2_\Gamma\right) \tag{1}$$

for some function $\ell(\cdot)$, where $\|X-\mu\|^2_\Gamma = (X-\mu)^T\Gamma^{-1}(X-\mu)$, $\mu \in \mathbb{R}^p$, and $\Gamma \in \mathbb{R}^{p\times p}$ is positive definite. For the standardized predictor $Z = \Gamma^{-1/2}(X-\mu)$, it has density $\ell(\|Z\|^2)$ with $\|Z\|^2 = Z^T Z$, which only depends on the length of Z. Because $Y \perp\!\!\!\perp Z|\eta^T Z$ and $Y \perp\!\!\!\perp X|(\Gamma^{-1/2}\eta)^T X$ imply each other, we may first find the Z-scaled central space $\mathcal{S}_{Y|Z}$ and then transform it back to the X-scale by $\mathcal{S}_{Y|X} = \Gamma^{-1/2}\mathcal{S}_{Y|Z}$.

Kernel inverse regression (KIR; [26]) suggests using kernel method to estimate the central space. It is an alternative to the popular method SIR, which estimates the inverse regression mean by slicing the response Y. [10] have shown that a single outlier can seriously distort the estimation of SIR. [21] demonstrated that SIR will fail when the distribution of X is elliptical with heavy tails. We suspect KIR will inherit these limitations. Our motivation is to propose robust procedures that perform as well as the classical KIR when the predictor X is multivariate normal, and keep up the good performances when X is contaminated by outliers or has distributions with heavy tails.

Three algorithms for robust KIR are proposed in this paper. One naive proposal is to use robust estimates of μ and Γ to standardize the predictor, and then implement robust PCA instead of PCA in the classical KIR algorithm. In our second proposal, by noticing that KIR essentially implements the local inverse mean, we suggest using the notion of multivariate local inverse median. Our third proposal suggests downweighting the effect of potential outliers. It is shown that while the improvement of the naive proposal over the classical KIR is limited, robust KIR can be effectively facilitated by either using a local inverse median or downweighing the outliers.

Through a newly defined sample influence function, we can detect influential points for the classical KIR estimation.

*Corresponding author.

Our one-step robust procedure is demonstrated to work as well as a two-step procedure, where we first detect and delete the influential points, and then perform classical KIR in the second step. Determination of the structural dimension d may also be distorted in the presence of outliers. One contribution is that our robust estimator of the the central space $\mathcal{S}_{Y|X}$ naturally leads to better estimation accuracy of the structural dimension.

The rest of the paper is organized as follows. In Section 2, we review the algorithm of KIR and propose three robust variations. Simulation studies and real data analysis are carried out in Sections 3 and 4 respectively. We conclude the paper with some discussions in Section 5.

2. KERNEL INVERSE REGRESSION AND ITS ROBUST VARIATIONS

We first briefly review the idea of kernel inverse regression. Given an i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, the location parameter μ and the dispersion parameter Γ in (1) can be estimated by $\hat{\mu} = \sum_{i=1}^n X_i/n$ and $\hat{\Gamma} = \sum_{i=1}^n (X_i - \hat{\mu})(X_i - \hat{\mu})^T/n$ respectively. Denote $\hat{Z}_i = \hat{\Gamma}^{-1/2}(X_i - \hat{\mu})$. For kernel function $K(\cdot)$ and bandwidth h, [26] proposed to estimate $E(Z|Y_j)$ by

(2)
$$\hat{E}_j = \sum_{i=1}^n \hat{Z}_i w_{ij}, \quad \text{where } w_{ij} = \frac{K[(Y_i - Y_j)h^{-1}]}{\sum_{i=1}^n K[(Y_i - Y_j)h^{-1}]}.$$

These local inverse mean estimates are stacked together to get $\hat{E} = \{\hat{E}_1, \ldots, \hat{E}_n\}$. The eigenvectors corresponding to the largest eigenvalues of $M_n = \hat{E}\hat{E}^T/n$ are then used to estimate the central space $\mathcal{S}_{Y|Z}$. Please refer to [26] for the $\sqrt{n}$-consistency of M_n. The estimate of $E(Z|Y_j)$ in (2) is a weighted average, and the observations with Y values closer to Y_j have larger weights. The estimate in (2) is also the solution to the following optimization problem:

$$\hat{E}_j = \underset{m \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \|\hat{Z}_i - m\|^2 w_{ij}, \tag{3}$$

where $\|\hat{Z}_i - m\|^2 = (\hat{Z}_i - m)^T(\hat{Z}_i - m)$.

2.1 A naive algorithm for robust kernel inverse regression

[10] studied the influence function of SIR and suggested that SIR may be seriously affected by a single outlier. KIR has the same limitation for the following reasons. First of all, using sample mean and sample variance to estimate μ and Γ is problematic with potential outliers. This can be addressed by using robust estimators of the location and dispersion parameters instead. A less obvious reason is that when Y_j has a large distance from all the other responses, the weights w_{ij} are very small for $i \neq j$ and will be dominated by w_{jj}. This means $E(Z|Y_j)$ is essentially estimated by $\hat{Z}_j$, which may be very large as X_j is a potential outlier. Thus the jth column of $\hat{E}$ may be unduly large, and will distort the eigenvalue decomposition of $M_n = \hat{E}\hat{E}^T/n$. To address this concern, we can use a robust version of PCA. The discussions above suggest the following naive robust KIR algorithm.

1. Find robust estimators $\hat{\mu}$ and $\hat{\Gamma}$ of the location and dispersion parameters in (1).
2. For $i = 1, \ldots, n$, calculate $\hat{Z}_i = \hat{\Gamma}^{-1/2}(X_i - \hat{\mu})$.
3. For $j = 1, \ldots, n$, calculate $\hat{E}_j = \sum_{i=1}^n \hat{Z}_i w_{ij}$ with w_{ij} defined in (2).
4. Estimate $E = \{E(Z|Y_1), \ldots, E(Z|Y_n)\}$ by $\hat{E} = \{\hat{E}_1, \ldots, \hat{E}_n\}$. Use robust PCA to calculate the eigenvectors $\hat{\eta}_1, \ldots, \hat{\eta}_d$ of $M_n = \hat{E}\hat{E}^T/n$, which correspond to the d largest eigenvalues of M_n.
5. Transform back to the X-scale central space and estimate $\mathcal{S}_{Y|X}$ with the column space of $\{\hat{\Gamma}^{-1/2}\hat{\eta}_1, \ldots, \hat{\Gamma}^{-1/2}\hat{\eta}_d\}$.

Many robust covariance estimators exist in the literature, among which the minimum covariance determinant estimator is one of the most popular. Please refer to [17] for a nice summary. For our purpose in step 1 above, we use the *covMcd* function from R package "robustbase". The implementation of *covMcd* uses the Fast MCD algorithm and the details can be found in [19]. For robust PCA in step 4, we use the *PCAgrid* function in R package "pcaPP", which computes robust principal components via projection pursuit [6, 11].

In step 4 above, we assume the structural dimension d of $\mathcal{S}_{Y|X}$ is known *a priori*. We will discuss how to estimate d in the numerical studies in Section 3. The estimator from the above algorithm will be referred to as KIR-R1.

2.2 A modified algorithm via local inverse median

In steps 1 and 4 of the algorithm in Section 2.1, we replace the original KIR algorithm with corresponding robust procedures. In step 3, however, the inverse mean $E(Z|Y)$ is still estimated by classical kernel method without adjusting for the effect of potential outliers. A natural idea here is to replace the inverse mean estimate with a robust location parameter estimate, such as the inverse median.

[9] proposed a robust version of the SIR algorithm, where they suggest replacing the intra slice mean with intra slice $L1$ median. Given $\hat{Z}_1, \ldots, \hat{Z}_n \in \mathbb{R}^p$, the $L1$ median [22] is defined as

$$\underset{m \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \|\hat{Z}_i - m\|.$$

It is obvious that if we replace the norm $\|\cdot\|$ in the above definition by the squared norm $\|\cdot\|^2$, the corresponding minimizer will be the sample mean of $\hat{Z}_i$. This leads us to consider replacing step 3 of the algorithm in Section 2.1 by

3′. For w_{ij} defined in (2), $j = 1, \cdots, n$, calculate

$$\hat{E}_j = \underset{m \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^{n} \|\hat{Z}_i - m\| w_{ij}. \tag{4}$$

If we replace the norm with the squared norm in (4), we will get (3) as described in Section 2.1. Instead of using the local inverse mean, we now use the local inverse median as the estimate of $E(Z|Y_j)$. As there is no explicit solution for (4), we use *optim* function in R to implement numerical minimization. We denote this second robust variation as KIR-R2.

Robust estimation of the location and scatter parameters in the multivariate case has been well-studied in the literature. For example, one can refer to [12]. Our estimator in (4) is different from the classical multivariate $L1$ median, and can be viewed as a localized $L1$ median. While $L1$ median is the point that minimizes the sum of the Euclidean distances to all points in the data set, (4) is minimizing a weighted sum, and data points with Y coordinates closer to Y_j have larger weights. When $p = 1$, our proposal (4) becomes a univariate local median regression, which is a special case of local quantile regression studied in [23].

2.3 A modified algorithm by downweighting potential outliers

[21] pointed out that when predictor X has elliptically-contoured distribution in (1), the inverse mean may not always exist. Contour projection was suggested in [21], where sliced inverse regression was implemented based on the weighted predictor. A similar idea of downweighting potential outliers was also considered under the setting of canonical correlations in [25]. Following these ideas, we modify (2) in Section 2.1 and define weights w^*_{ij}

$$w^*_{ij} = \frac{\|\hat{Z}_i\|^{-1} K[(Y_i - Y_j)h^{-1}]}{\sum_{i=1}^{n} \|\hat{Z}_i\|^{-1} K[(Y_i - Y_j)h^{-1}]}. \tag{5}$$

Step 3 of the algorithm in Section 2.1 is modified to be

3″. For $j = 1, \ldots, n$, calculate $\hat{E}_j = \sum_{i=1}^{n} \hat{Z}_i w^*_{ij}$ with w^*_{ij} defined in (5).

This modified algorithm will take into account the length of the standardized predictor, and observations further away from the center of the data cloud are given less weights. Our proposed weight is different from contour projection, which would perform classical KIR based on $\hat{Z}_i/\|\hat{Z}_i\|$. The resulting estimator from this modified algorithm will be denoted as KIR-R3.

3. SIMULATION STUDIES

Consider the following models:

$$\text{Model I}: \quad Y = \frac{X_1}{0.5 + (1.5 + X_2)^2} + .2\varepsilon,$$

$$\text{Model II}: \quad Y = (X_1 + 0.5)^3 + X_2 + .1\varepsilon,$$

where ε is standard normal independent of X. The central space for both models is then spanned by $\{e_1, e_2\}$, where $e_i \in \mathbb{R}^p$ is a vector with ith component 1, and 0 otherwise. Let $X = (X_1, \ldots, X_p)^T = W/\sqrt{\chi^2_\nu/\nu}$. Here $W \in \mathbb{R}^p$ is standard multivariate normal, χ^2_ν is a chi-squared distribution with ν degrees of freedom, and W is independent of χ^2_ν. Thus X follows a multivariate t distribution [13], which belongs to the elliptically-contoured distribution family in (1). We consider four scenarios for the distribution of X: (i) $\nu = \infty$, or X is multivariate normal; (ii) $\nu = 3$, or the predictor has a heavy tail with finite first moment; (iii) $\nu = 1$, or the predictor is multivariate Cauchy with no finite moments; (iv) $\nu = \infty$ with a single outlier, where we artificially distort the first observation in a multivariate normal sample and multiply it by 100.

Let β be the orthogonal basis of $\mathcal{S}_{Y|X}$ and $P_\beta = \beta(\beta^T\beta)^{-1}\beta^T$ be the orthogonal projection onto the column space of β. Denote $\hat{\beta}$ as an orthogonal estimate and $P_{\hat{\beta}}$ as its corresponding projection matrix. We follow [15] and measure the accuracy of the central space estimators by $\Delta = \|P_\beta - P_{\hat{\beta}}\|^2$. Smaller Δ implies a better estimator. To compare the performance of classical KIR with the robust proposals in Section 2, we summarize the results in Table 1 based on 100 repetitions. Fix $p = 4$ and consider sample sizes $n = 50, 100, 200$. Kernel method is not sensitive to the choice of density function $K(\cdot)$ but may be sensitive to the window width. We use Gaussian kernel and set $h = .1, .5, 1, 2$. Within each repetition, we choose the window width that corresponds to the smallest Δ for each method.

We make the following observations from Table 1. When X is multivariate normal in case (i), all three robust proposals have similar performances with classical KIR in Model I, and are slightly worse than KIR in Model II. In case (ii) when X is multivariate t with 3 degrees of freedom, all three robust methods improve over KIR. Such improvement becomes even more significant when X is Cauchy in case (iii). In case (iv), X is multivariate normal with a single outlier. KIR will fail, while the robust methods perform similarly to case (i) when X is normal with no outliers. Furthermore, in cases (i) and (ii), all four methods improve with increasing sample size. In cases (iii) and (iv), three robust procedures will keep improving as sample size increases. However, classical KIR is no longer consistent, and will not necessarily become better with larger sample sizes.

Among the three robust procedures, the naive algorithm in Section 2.1 has limited improvement over the classical KIR, and the modified algorithm in Section 2.3 based on downweighting outliers has the most significant overall improvement. Because the modified algorithm in Section 2.2 does not have an explicit solution for the minimization problem (4), numerical minimization is involved to calculate the local inverse median, which can be instable and time-consuming. This may explain why KIR-R2 is not as good as KIR-R3. To make this point clearer, we plot in Figure 1 the averages of $\Delta = \|P_\beta - P_{\hat{\beta}}\|^2$ based on 100 repetitions

Table 1. Averages and standard errors of $\Delta = \|P_\beta - P_{\hat\beta}\|^2$ based on 100 repetitions

Model	X	n	KIR	KIR-R1	KIR-R2	KIR-R3
I	(i)	50	.256(.026)	.205(.017)	.243(.021)	.193(.017)
		100	.119(.012)	.112(.010)	.145(.014)	.113(.011)
		200	.043(.004)	.048(.003)	.065(.006)	.050(.004)
	(ii)	50	.778(.055)	.459(.036)	.365(.031)	.304(.024)
		100	.702(.055)	.257(.023)	.205(.017)	.172(.015)
		200	.581(.057)	.126(.012)	.090(.007)	.065(.006)
	(iii)	50	1.583(.068)	1.311(.057)	.706(.046)	.550(.041)
		100	1.612(.054)	1.148(.065)	.372(.029)	.293(.021)
		200	1.823(.064)	1.188(.065)	.196(.021)	.144(.012)
	(iv)	50	1.062(.057)	.904(.059)	.289(.024)	.215(.016)
		100	1.057(.065)	.631(.050)	.145(.012)	.110(.009)
		200	.950(.059)	.457(.043)	.070(.005)	.052(.004)
II	(i)	50	.377(.040)	.510(.045)	.560(.047)	.510(.044)
		100	.138(.019)	.336(.035)	.411(.036)	.263(.026)
		200	.081(.011)	.282(.031)	.263(.025)	.230(.024)
	(ii)	50	.671(.057)	.810(.054)	.755(.051)	.597(.051)
		100	.590(.057)	.541(.043)	.457(.038)	.355(.035)
		200	.518(.051)	.312(.029)	.219(.023)	.185(.023)
	(iii)	50	.858(.065)	.901(.062)	.765(.054)	.766(.053)
		100	.886(.065)	.783(.055)	.519(.043)	.443(.040)
		200	.934(.066)	.534(.044)	.308(.029)	.293(.027)
	(iv)	50	.882(.068)	.558(.048)	.574(.049)	.505(.046)
		100	.986(.063)	.346(.028)	.393(.034)	.292(.032)
		200	1.036(.067)	.261(.031)	.246(.028)	.168(.015)

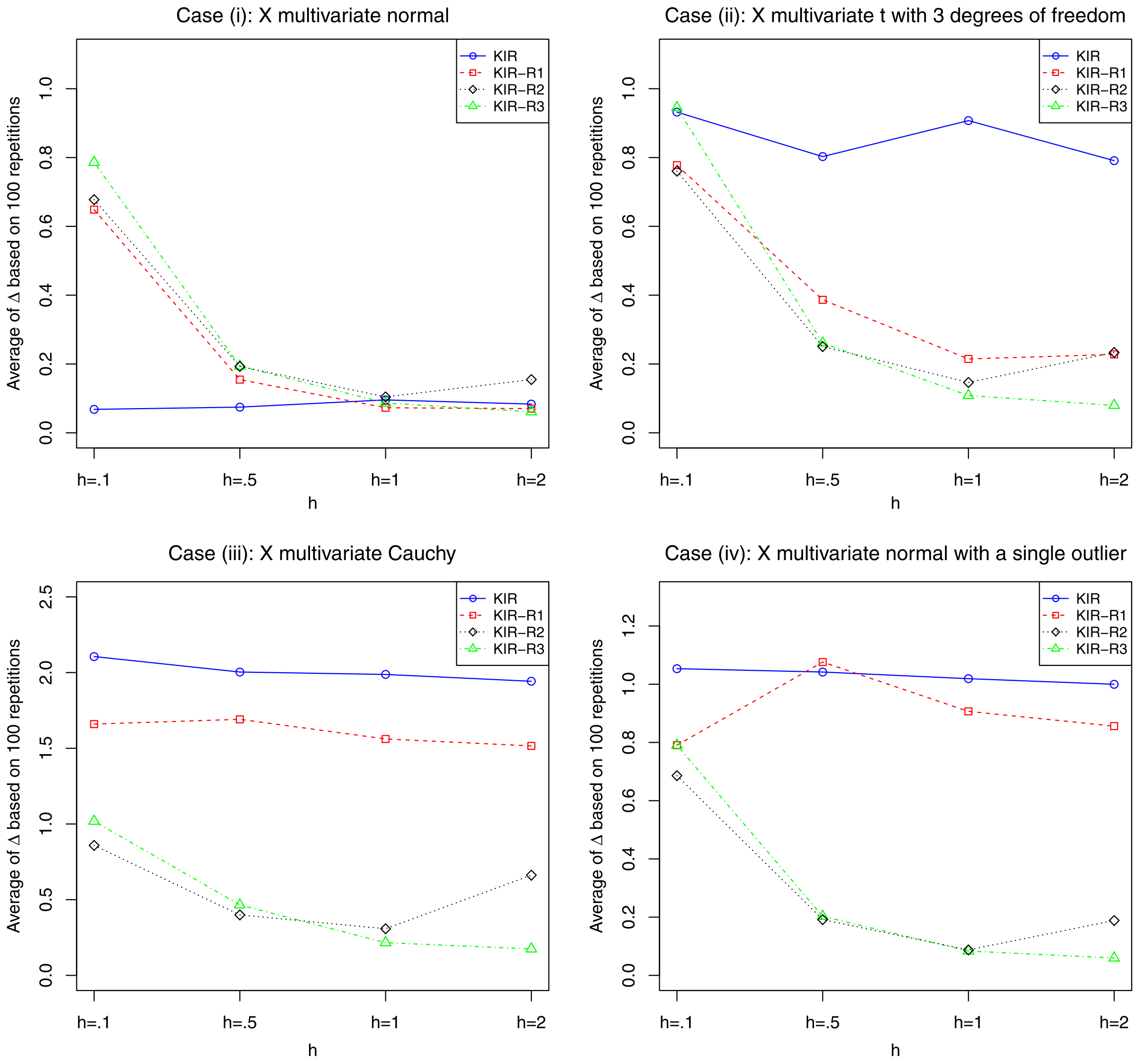
against window width $h = .1, .5, 1, 2$. The plots focus on Model I with $n = 200$. When X is normal, the robust procedures are comparable with KIR for h values other than $h = .1$. When X is contaminated by outliers or has distributions with heavy tails, the overall performances seem to be ordered from worst to best as: KIR, KIR-R1, KIR-R2, and KIR-R3.

Many methods exist in the SDR literature to determine the structural dimension d. Recall that we recover $\mathcal{S}_{Y|Z}$ by the column space of $\hat{E} = \{\hat{E}_1, \ldots, \hat{E}_n\}$. Thus estimating d is equivalent to estimate the number of nonzero eigenvalues of $M_n = \hat{E}\hat{E}^T/n$. Denote $\hat\lambda_1 \geq \cdots \geq \hat\lambda_p$ as the eigenvalues of M_n. A sequential test can be carried out as follows. For working structural dimension ℓ, test $H_0 : d = \ell$ versus $H_a : d > \ell$ for $\ell = 0, 1, \ldots, p-1$. Reject H_0 in favor of H_a if test statistic $\hat\Lambda_\ell = \sum_{i=\ell+1}^{p} \hat\lambda_i$ is larger than a certain threshold. The structural dimension is then estimated by $\hat{d} = \ell$ for the first ℓ such that H_0 is not rejected. The threshold used in [16] relies on the asymptotic distribution of $\hat\Lambda_\ell$, which depends on the normality of X and is not directly applicable in our case. We use a permutation test instead, which is free of the distribution of X. Please refer to [5] for details. For the ease of presentation, we focus on Model I with $n = 200$, and only compare classical KIR with KIR-R3.

The results of the permutation test based on 100 repetitions are summarized in Table 2. The proportions of correctly identifying $d = 2$ are highlighted in boldface. When X is multivariate normal in case (i), permutation test for KIR works well for all h. Permutation test for robust KIR works well except when $h = .1$. This agrees with what we observed from the first panel of Figure 1. When X has heavy tails in cases (ii) and (iii), or when X is contaminated by a single outlier in case (iv), permutation test for robust KIR does reasonably well for determining d with $h = 1$ or $h = 2$, and yields much better results than permutation test for classical KIR.

4. EMPIRICAL STUDIES

In this section, we analyze the 2004 automobile price data set, which can be downloaded from the Journal of Statistics Education data archive (*www.amstat.org/publications/jse/jse_data_archive.htm*). Due to the fact that KIR-R3 enjoys the best overall performance among all the robust proposals in the simulation studies, we will focus on comparisons between the classical KIR and KIR-R3. Originally, there are 428 cases, 16 predictor variables and there are some missing values. After removing the categorical variables and missing values from the original data, the remaining data contains $n = 387$ observations and $p = 8$ continuous predictors: Engine Size, Horsepower, City mpg, Highway mpg, Weight, Wheel Base, Length and Width. We standardize each predictor using its mean and standard deviation. The scatterplot matrix of the standardized predic-


Case (i): X multivariate normal

Case (ii): X multivariate t with 3 degrees of freedom

Case (iii): X multivariate Cauchy

Case (iv): X multivariate normal with a single outlier

Figure 1. Comparison of different h under Model I.

tors does not reveal strong violation of the elliptical distribution assumption.

To study the factors that affect the prices of automobiles, we take the manufacturer suggested retail price (MSRP) as the response. First we plot the histogram of the standardized MSRP in Figure 2. We see that the distribution of the response is highly skewed, suggesting some potential outliers exist in this data set. Next we use permutation tests to estimate the structural dimension d. For this data set, different choices of window width lead to consistent results in terms of estimating d. The permutation test based on KIR implies $d = 3$, while the test based on KIR-R3 suggests $d = 2$. As we have seen in Table 2, the permutation test based on KIR may be distorted when there are potential outliers. Thus we use $d = 2$ as the working structural dimension.

To estimate the central space, we use $h = .25$ for KIR and $h = 1.5$ for KIR-R3, which are chosen so that the corresponding estimators have the highest sample correlation with the response. These choices agree with previous findings in Figure 1 that KIR-R3 prefers larger h values. Denote the estimators from KIR and KIR-R3 as $\hat{\beta}_{\mathrm{KIR}} = (\hat{\beta}^{\mathrm{1st}}_{\mathrm{KIR}}, \hat{\beta}^{\mathrm{2nd}}_{\mathrm{KIR}})$, $\hat{\beta}_{\mathrm{KIR-R3}} = (\hat{\beta}^{\mathrm{1st}}_{\mathrm{KIR-R3}}, \hat{\beta}^{\mathrm{2nd}}_{\mathrm{KIR-R3}})$ re-

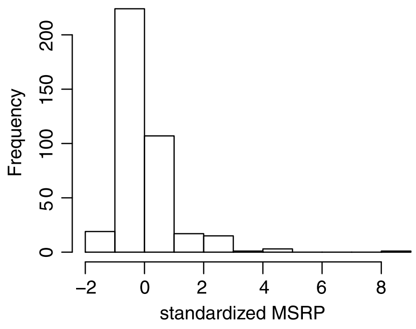


Figure 2. Histogram of standardized MSRP.

Table 2. Permutation test to estimate the structural dimension of Model I. Proportions based on 100 repetitions are reported

X		(i)		(ii)		(iii)		(iv)	
method		KIR	KIR-R3	KIR	KIR-R3	KIR	KIR-R3	KIR	KIR-R3
h=.1	$\hat{d}=0$	0	0	.03	0	.97	.12	.12	0
	$\hat{d}=1$	0	.56	.50	.75	.03	.82	.72	.48
	$\hat{d}=2$	**.92**	**.40**	**.28**	**.24**	**0**	**.06**	**.10**	**.44**
	$\hat{d}>2$	.08	.05	.19	.01	0	0	.06	.08
h=.5	$\hat{d}=0$	0	0	.02	.01	.94	.04	0	0
	$\hat{d}=1$	0	.03	.28	.16	.06	.59	.28	.02
	$\hat{d}=2$	**.91**	**.92**	**.41**	**.83**	**0**	**.37**	**.65**	**.90**
	$\hat{d}>2$	.09	.05	.29	0	0	0	.07	.08
h=1	$\hat{d}=0$	0	0	.03	.01	.88	.04	0	0
	$\hat{d}=1$	0	0	.24	.03	.12	.33	.21	0
	$\hat{d}=2$	**.94**	**.98**	**.45**	**.94**	**0**	**.62**	**.68**	**.95**
	$\hat{d}>2$	.06	.02	.28	.01	0	.01	.11	.05
h=2	$\hat{d}=0$	0	0	.01	.01	.89	.03	0	0
	$\hat{d}=1$	0	0	.09	0	.10	.25	.14	0
	$\hat{d}=2$	**.95**	**.93**	**.59**	**.96**	**.01**	**.69**	**.74**	**.96**
	$\hat{d}>2$	.05	.07	.31	.03	0	.03	.12	.04

Table 3. Coefficients estimation of the automobile price data

X	Engine Size	Horsepower	City mpg	Highway mpg	Weight	Wheel Base	Length	Width
$\hat{\beta}_{\text{KIR}}^{\text{1st}}$	.009	.828	$-.077$	.288	.403	$-.170$	$-.018$	$-.183$
$\hat{\beta}_{\text{KIR}}^{\text{2nd}}$	.095	.281	.572	$-.328$	$-.393$	$-.483$	.290	.074
$\hat{\beta}_{\text{KIR-R3}}^{\text{1st}}$	$-.198$	.588	$-.411$	.298	.483	.230	$-.183$	$-.196$
$\hat{\beta}_{\text{KIR-R3}}^{\text{2nd}}$	.324	.362	.788	$-.099$	.234	.093	$-.143$	$-.223$

spectively, and we summarize them in Table 3. Each row of Table 3 provides the coefficients of the predictors, and is used to create a score. For example, the first KIR score is

$$\begin{aligned} X^T\hat{\beta}_{\text{KIR}}^{\text{1st}} &= .009\times\text{Engine Size} + .828\times\text{Horsepower} \\ &\quad - .077\times\text{City mpg} + .288\times\text{Highway mpg} + .403\times\text{Weight} \\ &\quad - .170\times\text{Wheel Base} - .018\times\text{Length} - .183\times\text{Width}. \end{aligned}$$

The dominating factors for the above score is Horsepower and Weight with positive coefficients. Its sample correlation with the response is $corr(X^T\hat{\beta}_{\text{KIR}}^{\text{1st}}, Y) = .857$, which agrees with the intuition that larger Horsepower and larger Weight correspond to more expensive cars. In Figure 3, we plot the price against $X^T\hat{\beta}_{\text{KIR}}^{\text{1st}}$ and clearly see an increasing trend. The plot of the price versus the first KIR-R3 score reveals a similar trend. Actually, $X^T\hat{\beta}_{\text{KIR}}^{\text{1st}}$ and $X^T\hat{\beta}_{\text{KIR-R3}}^{\text{1st}}$ is highly correlated with sample correlation .891. On the other hand, the second scores of KIR and KIR-R3 turn out to be somewhat different with $corr(X^T\hat{\beta}_{\text{KIR}}^{\text{2nd}}, X^T\hat{\beta}_{\text{KIR-R3}}^{\text{2nd}}) = .364$.

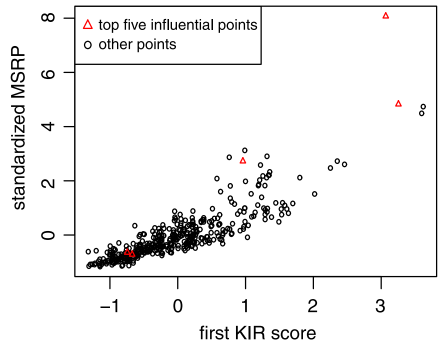


Figure 3. Sufficient plot of Y v.s. $X^T\hat{\beta}_{\text{KIR}}^{\text{1st}}$.

We use the following scheme to compare the seemingly different KIR and KIR-R3 estimators. The full data set is randomly split into two subsets with approximately equal sample sizes ($n_1 = 193$ and $n_2 = 194$ to be exact). Denote $\hat{\beta}_{\text{KIR}}^{(n_1)}$ as the estimator of KIR based on the first subset, and $\hat{\beta}_{\text{KIR}}^{(n_2)}$ is based on the second subset. Then we calculate

$$\delta_{\text{KIR}} = \left\| P_{\hat{\beta}_{\text{KIR}}^{(n_1)}} - P_{\hat{\beta}_{\text{KIR}}^{(n_2)}} \right\|^2, \tag{6}$$

which measures the difference between the two subset estimators. Ideally, this distance should be small as the two subsets are from the same data source. The difference $\delta_{\text{KIR-R3}}$ is calculated in a parallel fashion. Based on 100 repetitions, the average of δ_{KIR} is 1.243, and the average of $\delta_{\text{KIR-R3}}$ improves to .725. The estimator based on KIR-R3 is more robust, and thus corresponds to smaller differences between the subsets.

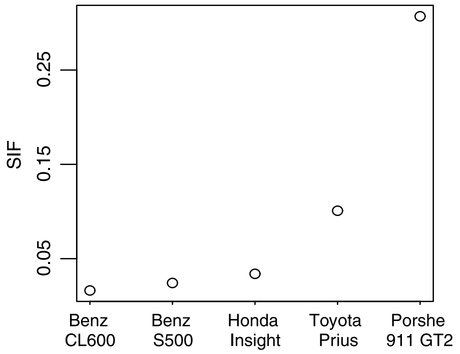


Figure 4. SIF values versus the top five influential points.

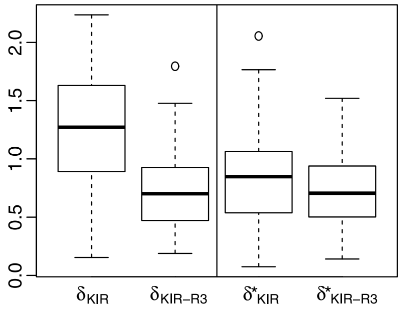


Figure 5. δ and δ^* based on 100 repetitions. Left panel: all complete cases with $n = 387$. Right panel: $n^* = 382$ after deleting the top five influential points.

Following the suggestions of an anonymous referee, we now study the sample influence function of KIR. Recall that $\hat{\beta}_{\rm KIR}$ denotes the KIR estimator based on the full data set. Denote $\hat{\beta}_{(i),\rm KIR}$ as the KIR estimator based on the data with the ith observation deleted. We define the sample influence function of the ith point as follows

$$SIF(i) = \left\| P_{\hat{\beta}_{\rm KIR}} - P_{\hat{\beta}_{\rm (i),KIR}} \right\|^2. \tag{7}$$

This definition is similar to Cook's distance in the regression setting [4], and it measures the effect of deleting a given observation. Next we use SIF to detect the influential points in the automobile data set. After ordering the observations by their SIF values, we plot the SIF for the top five influential points in Figure 4. The SIF values of the remaining points are all below .015, and are thus excluded. It is not really surprising that these particular points turn out to be influential. Benz CL600, Benz S500 and Porshe 911 GT2 are all high-end luxury cars. Honda Insight and Toyota Prius are both electric cars and are extremely fuel efficient. These five most influential points are also highlighted in Figure 3 for easy visualization.

We have seen that KIR-R3 is more robust than KIR in terms of the difference measure (6). To get a better understanding about how KIR is affected by a few influential points, we delete the five most influential points, and recalculate $\delta^*_{\rm KIR}$ and $\delta^*_{\rm KIR-R3}$ based on the remaining $n^* = 382$ observations. The boxplots of $\delta_{\rm KIR}$, $\delta_{\rm KIR-R3}$, $\delta^*_{\rm KIR}$ and $\delta^*_{\rm KIR-R3}$ are summarized in Figure 5. With $n = 387$ complete observations, we see from the left panel that KIR-R3 is significantly better than KIR. The right panel represents a common two-step procedure, where we detect and delete the outliers in the first step, and then carry out dimension reduction subsequently. This is a viable strategy as KIR greatly improves. We clearly see from Figure 5 that KIR is sensitive to the influential points. On the other hand, KIR-R3 performs as well as KIR when the outliers are deleted, and is much better in the presence of potential outliers.

5. DISCUSSIONS

In this paper, we discuss robust variations of kernel inverse regression for sufficient dimension reduction. Our robust proposals work as well as classical KIR when the predictor is multivariate normal, and become significantly better when X has heavy-tailed distributions or when X is contaminated by outliers. Both local inverse median and weighted KIR are demonstrated to be effective. Because local inverse median involves heavy computation and may not be stable when p is large, we prefer weighted KIR. Our experience indicates that weighted KIR works very well in both the simulation setting and the empirical studies. Better accuracy of estimating $\mathcal{S}_{Y|X}$ can lead to better accuracy of estimating d. Our robust estimators together with the permutation test can estimate the structural dimension effectively in the presence of potential outliers.

[2] suggested that predictor contributions can be tested without knowing the link function under the SDR framework, extensions of which have been studied in [20, 24]. When X is heavy-tailed or contaminated by outliers, tests based on classical SDR methods will be likely to fail. Development of tests based on robust procedures is warranted. We use sample influence function (7) to detect influential points in the classical KIR estimation. Population level influence function as well as break-down point properties of sliced inverse regression have been studied in [10, 18], and the corresponding development for kernel inverse regression is currently under investigation.

ACKNOWLEDGEMENT

This research was supported in part by the U. S. National Science Foundation Grant DMS-1106577 awarded to the first author. We sincerely thank two anonymous referees for their helpful comments that led to a much-improved paper.

Received 15 December 2011

REFERENCES

[1] Cook, R. D. (1998). *Regression Graphics*. New York: Wiley. MR1645673

[2] Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *The Annals of Statistics* **32** 1062–1092. MR2065198

[3] Cook, R. D. and Nachtsheim, C. J. (1994). Reweighting to achieve elliptically contoured covariates in regression. *Journal of the American Statistical Association* **89** 592–599.

[4] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall. MR0675263

[5] Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian and New Zealand Journal of Statistics* **43** 147–199. MR1839361

[6] Croux, C., Filzmoser, P. and Oliveira, M. R. (2007). Algorithms for projection-pursuit robust principal component analysis. *Chemometrics and Intelligent Laboratory Systems* **87** 218–225.

[7] Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: Second-order methods. *Biometrika* **97** 279–294. MR2650738

[8] Eaton, M. L. (1986). A characterization of spherical distributions. *Journal of Multivariate Analysis* **34** 439–446. MR0866075

[9] Gather, U., Hilker, T. and Becker, C. (2001). A Robustified Version of Sliced Inverse Regression. *Statistics in Genetics and in the Environmental Sciences* (L. T. Fernholz, S. Morgenthaler & W. Stahel, eds.). Basel: Birkhäuser, 147–157. MR1843175

[10] Gather, U., Hilker, T. and Becker, C. (2002). A note on outlier sensitivity of sliced inverse regression. *Statistics* **36** 271–281. MR1923466

[11] Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics* **13**(2) 435–475. MR0790553

[12] Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. New Jersey: John Wiley & Sons, Inc. MR2488795

[13] Lange, K. L., Little, R. J. A. and Taylor, J. M. G. (1989). Robust statistical modeling using t distribution. *Journal of the American Statistical Association* **84** 881–896. MR1134486

[14] Li, B. and Dong, Y. (2009). Dimension reduction for non-elliptically distributed predictors. *The Annals of Statistics* **37** 1272–1298. MR2509074

[15] Li, B., Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *The Annals of Statistics* **33** 1580–1616. MR2166556

[16] Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86** 316–342. MR1137117

[17] Maronna, R. and Zamar, R. (2002). Robust estimation of location and dispersion for high dimensional datasets. *Technometrics* **44** 307–317. MR1939680

[18] Prendergast, L. A. (2005). Influence functions for sliced inverse regression. *Scand. J. Statist.* **32** 385–404. MR2204626

[19] Rousseeuw, P. J. and van Driessen K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.

[20] Shao, Y., Cook, R. D. and Weisberg, S. (2007). Marginal tests with sliced average variance estimation. *Biometrika* **94** 285–296. MR2331487

[21] Wang, H., Ni, L. and Tsai, C. L. (2008). Improving dimension reduction via contour-projection. *Statistica Sinica* **18** 299–311. MR2416908

[22] Weber, A. (1909). *Uber Den Standard Der Industrien, Tubingen*. English translation by C. J. Freidrich (1929). Alfred Weber's Theory of Location of Industries. Chicago: Chicago University Press.

[23] Yu, K. and Jones, M. C. (1998). Local linear quantile regression. *Journal of the American Statistical Association* **93** 228–237. MR1614628

[24] Yu, Z., Dong, Y. and Fang, Y. (2010). Marginal coordinate tests for central mean subspace with principal Hessian directions. *Chinese Journal of Applied Probability and Statistics* **26** 544–552. MR2779512

[25] Zhou, J. (2009). Robust dimension reduction based on canonical correlation. *Journal of Multivariate Analysis* **100** 195–209. MR2460487

[26] Zhu, L. X. and Fang, K. T. (1996). Asymptotics for kernel estimate of sliced inverse regression. *The Annals of Statistics* **3** 1053–1068. MR1401836

Yuexiao Dong
Temple University
Philadelphia, PA, 19122
USA
E-mail address: ydong@temple.edu

Zhou Yu
East China Normal University
Shanghai, 200241
P.R. China
E-mail address: zyu@stat.ecnu.edu.cn

Yizhi Sun
Temple University
Philadelphia, PA, 19122
USA
E-mail address: tuc33565@temple.edu