Web Appendix:
Robust Recovery of the Central Subspace for Regression Using the Influence Function of the Rényi Divergence

Ross Iaci and T.N. Sriram

B Heuristic argument for robustness

Let $\widehat{\mathbf{A}}_{\alpha_l}$ denote the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$ for a fixed level of the tuning parameter, $\alpha_l \in (0,1)$. The monotonicity property of $\mathcal{R}_\alpha(\mathbf{A})$ discussed in Section 2.2, together with article-(4) and article-(5), suggests that for any two values of the tuning parameter α_1 and α_2 such that $\alpha_1 < \alpha_2$, $\widehat{\mathcal{R}}_{\alpha_1}(\widehat{\mathbf{A}}_{\alpha_1}) \leq \widehat{\mathcal{R}}_{\alpha_2}(\widehat{\mathbf{A}}_{\alpha_2}) \leq \widehat{\mathcal{D}}_{KL}(\widehat{\mathbf{A}}_1)$. Note that, $\mathcal{D}_{KL}(\mathbf{A}_1) \leq \mathcal{D}_{KL}(\mathbf{I})$ holds by Proposition 3 part (ii) of Yin and Cook (2005), and that the KL based method is fundamentally a likelihood procedure, implying that $\widehat{\mathbf{A}}_1$ would be asymptotically more efficient than $\widehat{\mathbf{A}}_{\alpha_l}$; however, $\widehat{\mathbf{A}}_1$ is more sensitive to the presence of extreme observations. This is discussed in the following heuristic argument to illustrate why the estimator in article-(4) possesses robustness against data contamination.

For n fixed, let $k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right) = \widehat{f}(y_i, \mathbf{A}^\top \mathbf{x}_i)/\{\widehat{f}(y_i)\widehat{f}(\mathbf{A}^\top \mathbf{x}_i)\}$ and view $\widehat{\mathcal{R}}_\alpha(\mathbf{A})$ as a function of α, then taking the limit using L'Hospital's rule,

$$\begin{aligned}
\lim_{\alpha\to 1} \widehat{\mathcal{R}}_\alpha(\mathbf{A}) &= \lim_{\alpha\to 1} \frac{1}{\alpha - 1} \ln\left[\frac{1}{n}\sum_{i=1}^{n} \left\{k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right)\right\}^{\alpha-1}\right] \\
&= \lim_{\alpha\to 1} \frac{\partial}{\partial \alpha}\left\{ \ln\left[\frac{1}{n}\sum_{1=1}^{n} \left\{k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right)\right\}^{\alpha-1}\right]\right\} \\
&= \lim_{\alpha\to 1} \frac{1}{n}\sum_{1=1}^{n} \left\{k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right)\right\}^{\alpha-1} \ln\left\{k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right)\right\} \Bigg/ \frac{1}{n}\sum_{i=1}^{n} \left\{k\left(y_i, \mathbf{A}^\top \mathbf{x}_i\right)\right\}^{\alpha-1}.
\end{aligned}$$

This implies that for α close to 1

$$\widehat{\mathcal{R}}_\alpha(\mathbf{A}) \approx \sum_{i=1}^{n} \widehat{w}_{(i,\alpha)} \ln\left\{\frac{\widehat{f}(y_i, \mathbf{A}^\top \mathbf{x}_i)}{\widehat{f}(y_i)\widehat{f}(\mathbf{A}^\top \mathbf{x}_i)}\right\}, \tag{B.1}$$

which is a weighted version of the sample index $\widehat{\mathcal{D}}_{KL}(\mathbf{A})$ of Yin and Cook (2005), with weights

$$\widehat{w}_{(i,\alpha)} = \left\{\frac{\widehat{f}(y_i, \mathbf{A}^\top \mathbf{x}_i)}{\widehat{f}(y_i)\widehat{f}(\mathbf{A}^\top \mathbf{x}_i)}\right\}^{\alpha-1} \Bigg/ \sum_{j=1}^{n} \left\{\frac{\widehat{f}(y_j, \mathbf{A}^\top \mathbf{x}_j)}{\widehat{f}(y_j)\widehat{f}(\mathbf{A}^\top \mathbf{x}_j)}\right\}^{\alpha-1}.$$

Therefore, for values of $\alpha \in (0,1)$, (B.1) naturally down-weighs any outlying observation, say $(y_{i^*}, \mathbf{x}_{i^*})$, since $\widehat{w}_{(i^*,\alpha)} < 1$. However, at $\alpha = 1$ the sample estimate is $\widehat{\mathcal{D}}_{KL}(\mathbf{A})$ and thus, all observations, including outliers, have weights $\widehat{w}_{(i,1)} = 1$. This implies that by selecting α close to 1, all weights approach 1, improving the asymptotic efficiency of the estimated coefficient matrix $\widehat{\mathbf{A}}$. Therefore, $\widehat{\mathcal{R}}_\alpha(\mathbf{A})$ offers a compromise between efficiency and robustness, with the extent of the concession controlled by the level of the tuning parameter α. For this reason, the selection of an optimal value of α, and a more formal assessment of robustness, is necessary and is accomplish through the study of influence functions.

C Simulation studies

C.1 Introduction

In this section, various regression models with differing levels of asymmetric contamination are investigated to study the robustness of our method in estimating a basis for $\mathcal{S}_{Y|\mathbf{X}}$. For all simulations, the level of the tuning parameter is taken on the grid $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. Also, as discussed in Section 2.6, all calculations are performed in the whitened scale and therefore, the orthonormal constraints are assumed for both $\widehat{\mathbf{A}}$ and $\mathbf{A}$. However, for the ease in exposition the notation Y and $\mathbf{X}$ is maintained, with the regression models described in the original scale and the estimated coefficient matrix $\widehat{\mathbf{A}}$ compared to the normed matrix $\Sigma^{1/2}\mathbf{A}$. Three different studies, for a total of five simulations, are used to examine the performance our methodology in the presence of outliers. In addition, the scale of the SIF values are not important and therefore, the results are reported in the *standardized* scale $(n-1)^{-1}\text{SIF}\big(\rho_{BC}, \widehat{F}, \mathbf{w}_i\big) = \big\{\rho_{BC}(\widehat{\mathbf{A}}_{(i)}, \widehat{\mathbf{A}}) - 1\big\}$.

Two measures between the estimated and true coefficient matrices $\widehat{\mathbf{A}}$ and $\mathbf{A}$ are used to quantify the accuracy of the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$. The definitions of the matrix 2-norm and determinant, and related results, are the same as those given in Section 4.1.

The first accuracy measure is an L_2 norm distance between $P_{\mathcal{S}(\mathbf{A})}$ and $P_{\mathcal{S}(\widehat{\mathbf{A}})}$, defined as

$$L_{2(D)}\big(\widehat{\mathbf{A}}, \mathbf{A}\big) = \big|\big|\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top - \mathbf{A}\mathbf{A}^\top\big|\big|_2 = \big|\big|P_{\mathcal{S}(\mathbf{A})} - P_{\mathcal{S}(\widehat{\mathbf{A}})}\big|\big|_2. \tag{C.1}$$

Importantly, note that $0 \leq L_{2(D)} \leq 2$, since $\big|\big|P_{\mathcal{S}(\mathbf{A})} - P_{\mathcal{S}(\widehat{\mathbf{A}})}\big|\big|_2 \leq \big|\big|P_{\mathcal{S}(\mathbf{A})}\big|\big|_2 + \big|\big|(-1)P_{\mathcal{S}(\widehat{\mathbf{A}})}\big|\big|_2$ $= \big|\big|P_{\mathcal{S}(\mathbf{A})}\big|\big|_2 + \big|\big|P_{\mathcal{S}(\widehat{\mathbf{A}})}\big|\big|_2 \leq 1 + 1 = 2$.

The second measure of accuracy is the correlation between $\mathcal{S}(\mathbf{A})$ and $\mathcal{S}(\widehat{\mathbf{A}})$ using the square root of Hotelling's (1936) squared vector correlation coefficient,

$$\rho_{HC}\big(\widehat{\mathbf{A}}, \mathbf{A}\big) = \sqrt{\big|(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^T\widehat{\mathbf{A}}(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}})^{-1}\widehat{\mathbf{A}}^\top\mathbf{A}\big|} = \sqrt{\big|\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}\big|} = \Big(\prod_{i=1}^{\mathrm{d}}\lambda_i\Big)^{\frac{1}{2}}. \tag{C.2}$$

As in Section 4.2, the λ_i are the eigenvalues of $\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}$, $0 \leq \rho_{HC}\big(\widehat{\mathbf{A}}, \mathbf{A}\big) \leq 1$, $\rho_{HC}\big(\widehat{\mathbf{A}}, \mathbf{A}\big) = 1$ implies that $\mathcal{S}(\mathbf{A})$=$\mathcal{S}(\widehat{\mathbf{A}})$, and $\rho_{HC}\big(\widehat{\mathbf{A}}, \mathbf{A}\big) = 0$ when the subspaces are orthogonal.

C.2 Simulation methodology and regression models

Uncontaminated error terms in the regression models are generated from a $N(0, \sigma)$ distribution, while asymmetric outlying observations are generated at random from a uniform distribution on the interval $(0, \theta)$ with probability $(1 - \pi)$, $\pi \in \{.95, .90\}$. In the study descriptions below, this is denoted as $\varepsilon \sim N(0, \sigma)\mathcal{I}(\pi) + U(0, \theta)\{1 - \mathcal{I}(\pi)\}$, where $\mathcal{I}(\pi) = 1$ with probability π and 0 with probability $(1 - \pi)$.

The distributions of the predictor variables $\mathbf{X} = (X_1, \ldots, X_{10})^\top$ and model error terms of each study are summarized as follows:

Study 1: $\mathbf{X}_{10} \sim N(\mathbf{0}, \mathbf{I})$; $\varepsilon \sim N(0, \sigma = 0.5)\mathcal{I}(\pi) + U(0, 50)\{1 - \mathcal{I}(\pi)\}$, $\pi \in \{.95, .90\}$.

Study 2: $X_1 \sim t(25)$, $X_2, X_3 \sim t(5)$, $X_4, X_5 \sim N(0, 1)$, $X_6 \sim \Gamma(4, 1)$, $X_7 \sim N(0, 1)$, $X_8 \sim \chi^2_{(3)}$, $X_9 \sim \Gamma(3, 2)$, $X_{10} \sim N(0, 1)$; $\varepsilon \sim N(0, \sigma = .3)\mathcal{I}(\pi) + U(0, 20)\{1 - \mathcal{I}(\pi)\}$, $\pi \in \{.95, .90\}$.

Study 3: $X_1 \sim \Gamma(4, 3)$, $X_2 \sim t(15)$, $X_3 \sim N(0, 1)$, $X_4 \sim \chi^2_{(3)}$, $X_5 \sim t(20)$, $X_6 \sim t(25)$, $X_7 \sim N(0, 1)$, $X_8 \sim \Gamma(10, 2)$, $X_9 \sim \chi^2_{(6)}$, $X_{10} \sim N(0, 1)$; $\varepsilon \sim N(0, \sigma = .3)\mathcal{I}(\pi) + U(0, 20)\{1 - \mathcal{I}(\pi)\}$, $\pi \in \{.95, .90\}$.

The regression models for each of the simulations for studies 1 - 3 are summarized in Table 1.

Simulation	Model	True Coefficient Matrices
	Study 1	
I	$Y = \mathbf{A}^\top\mathbf{X} + \varepsilon$	$\mathbf{A} = (1, 2, 0, 0, 0, \ldots, 0)^\top$
II	$Y = \mathbf{A}^\top\mathbf{X} + \varepsilon$	$\mathbf{A} = (1, 1, 1, 1, 0, \ldots, 0)^\top$
III	$Y = (\mathbf{A}^\top\mathbf{X})^2 + \varepsilon$	$\mathbf{A} = (1, 2, 3, 0, 0, \ldots, 0)^\top$
	Study 2	
I	$Y = \mathbf{a}_1^\top\mathbf{X}\left(\mathbf{a}_2^\top\mathbf{X} + 1\right) + \varepsilon$	$\mathbf{A} = [(1, 0, \ldots, 0)^\top; (0, 1, 0, \ldots, 0)^\top]$
	Study 3	
I	$Y = \frac{\mathbf{a}_1^\top\mathbf{X}}{0.5 + \left(\mathbf{a}_2^\top\mathbf{X} + 1.5\right)^2} + \varepsilon$	$\mathbf{A} = [(1, 0, \ldots, 0)^\top; (0, 1, 0, \ldots, 0)^\top]$

Table 1: Simulation regression models.

Note that, Study 3 considers a model that was used in Prendergast (2006) to illustrate their methods ability to detect influential observations using SIR, but not necessarily to examine the robustness of the procedure. Different from their numerical study, the predictors are not all normal, but complicated almost entirely with variables that follow a variety of skewed and heavy-tailed distributions, which are then contaminated with errors from a $U(0, 20)$ distribution. A randomly selected simulated dataset from each of Studies 2 and 3 is plotted in Figure 1 to illustrate each

type of regression relationship and importantly, the effects on these nonlinear associations due to contamination.

For a dataset created according to the above specifications an estimate of the coefficient matrix $\widehat{\mathbf{A}}$ is calculated and compared to the true basis $\mathbf{A}$ using (C.1) and (C.2). This process is repeated $n_s = 500$ times and the overall accuracy in estimating $\mathbf{A}$ quantified by taking the averages $\overline{L}_{2(D)} = \frac{1}{n_s}\sum_{j=1}^{n_s} L_{2(D)}(\widehat{\mathbf{A}}^j, \mathbf{A})$ and $\overline{\rho}_{HC} = \frac{1}{n_s}\sum_{j=1}^{n_s} \rho_{HC}(\widehat{\mathbf{A}}^j, \mathbf{A})$, where $\widehat{\mathbf{A}}^j$ is the estimated coefficient matrix for the j^{th} simulated dataset. Note that, for all simulations the standard errors of the means for both measures are less than 10^{-1} and not reported in the tables for brevity.

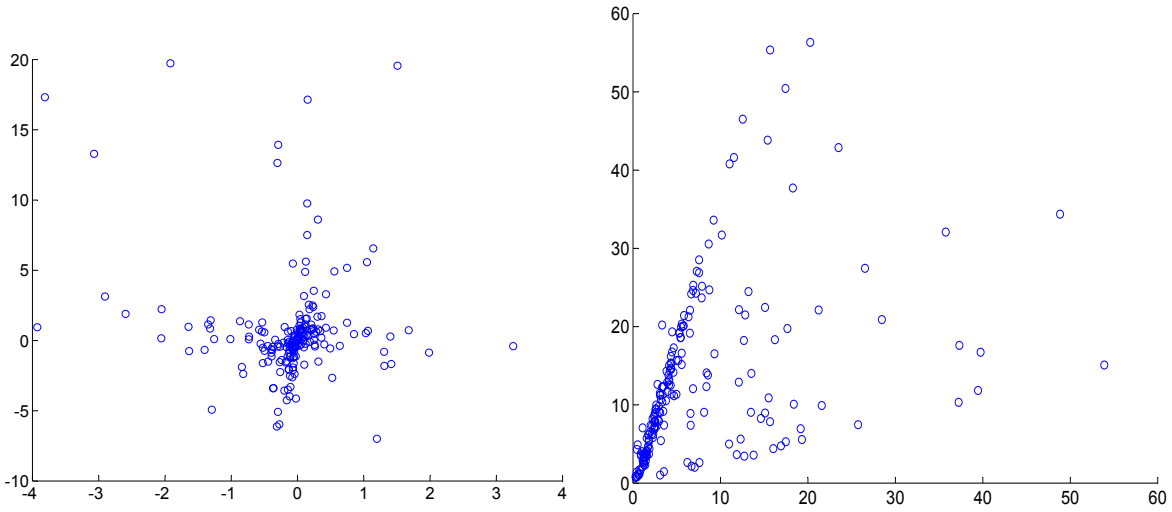

Figure 1: $n = 200$, $\pi = .95$. Data y versus $\mathbf{A}^\top\mathbf{x}$ plots, Simulation I. Left panel: Study 2, y versus $\mathbf{a}_1^\top\mathbf{x}(\mathbf{a}_2^\top\mathbf{x} + 1)$. Right panel: Study 3, y versus $\mathbf{a}_1^\top\mathbf{x}/\{0.5 + (\mathbf{a}_2^\top\mathbf{x} + 1.5)^2\}$.

C.3 Results Simulation Study 1

For all simulations of Study 1, the accuracy results between the actual and estimated bases are reported in Table 2 at each of the two levels of contamination, sample sizes $n = 200, 300$, and for brevity, only three values of the tuning parameter $\alpha = 0.2, 0.5$ and 0.8. Also, for succinctness the results for $n = 200$ are discussed below, with $n = 300$ numerically illustrating the theoretical consistency result in Section 2.5.

For the linear regression models in Simulations I and II, the mean correlations $\overline{\rho}_{HC}$ between the estimated and true basis $\mathbf{A}$ of $\mathcal{S}_{Y|\mathbf{X}}$ are all larger than .98 for both contamination levels and analogously, the mean $\overline{L}_{2(D)}$ distances are smaller with lower contamination. Since the L_2 distance in (C.1) takes on values on the interval $[0, 2]$, the $\overline{L}_{2(D)}$ values show a more definitive change in value moving from lower to higher levels of contamination and thus, can be viewed as a more sensitive accuracy measure.

Next, a dataset of size $n = 200$ was selected from this simulation to illustrate the methods discussed in Section 4. The SIF based AUC_α values, $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, of Section 4.4 are given in the plots in Figure 2 with the lowest area corresponding to $\alpha = 0.1$ and thus, the level of the tuning parameter that provides the most robust index. Next, Figure 3 shows that all the structural dimension detection methods discussed in Section 4 correctly estimate the dimension $\widehat{d} = 1$, with the boxplots of the bootstrap $1 - \rho_{HC}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values article-(13) and the |SIF| values $|s_{(i,k)}|$, $k = 1, 2, \ldots, 10$, having the smallest centers with the least variability. However, in addition to the capacity to estimate d, the SIF based methods can also be used to identify the most robust level of α and thereby, provide a foundation for more comprehensive methods with less computational effort.

The results in Table 2 for the linear relationship involving additional predictors, Simulation II, and the quadratic functional association between the response and disproportionately weighted predictor variables in Simulation III, are analogous and therefore, show that our method accurately recovers both linear and nonlinear regression DR directions in the presence of high contamination.

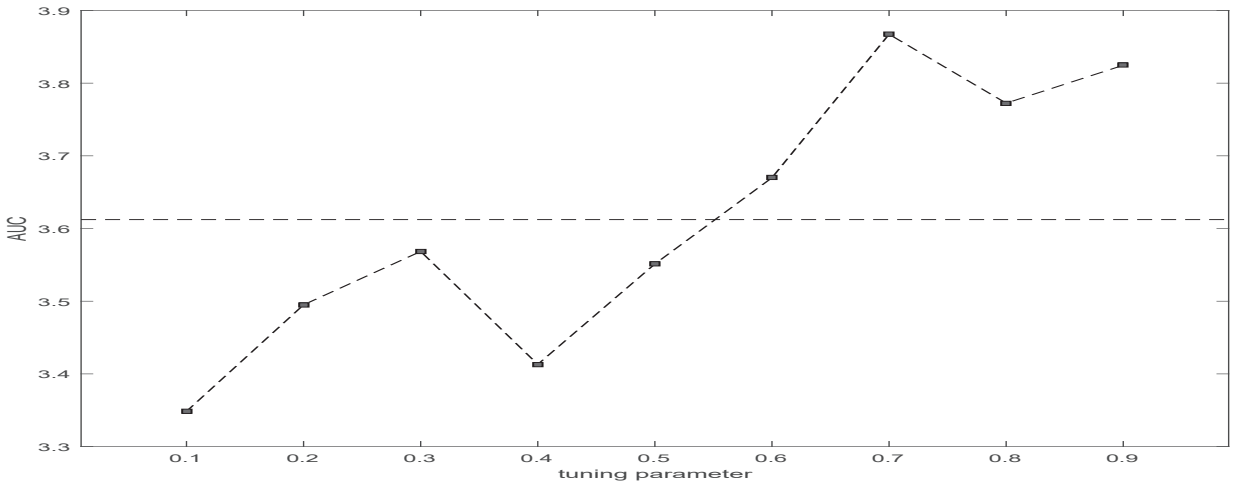

Figure 2: $n = 200, \pi = .90$, $\alpha = 0.1$. AUC_α values, $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$. (Study 1 Simulation I).

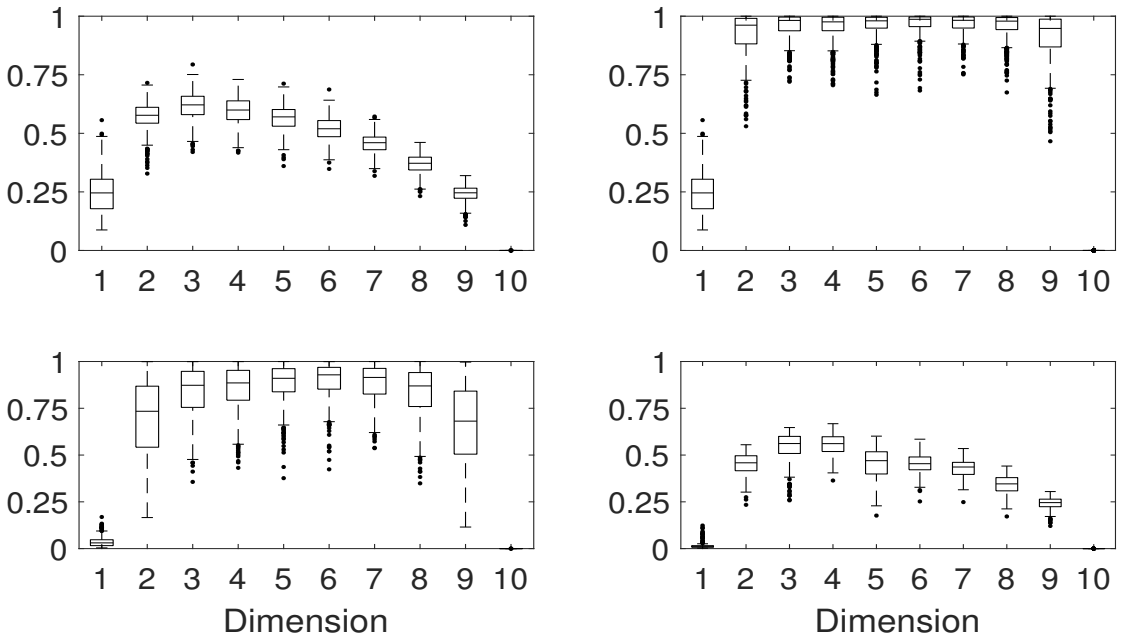

Figure 3: $n = 200, \pi = .90$, $\alpha = 0.1$. Dimension boxplots $k = 1, 2, \ldots, 10$. Top left panel: boxplots of bootstrap $\rho_{BC^*}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values. Top right panel: boxplots of bootstrap $L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values. Bottom left panel: boxplots of $1 - \rho_{HC}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values. Bottom right panel: boxplots of |SIF| values $|s_{(i,k)}|$. (Study 1 Simulation I).

Study 1									
	Simulation I			**Simulation II**			**Simulation III**		
α	0.2	0.5	0.8	0.2	0.5	0.8	0.2	0.5	0.8
$n = 200$									
					$\pi = .95$				
$\overline{\rho}_{HC}$	.9965	.9967	.9965	.9958	.9957	.9952	.9978	.9986	.9989
$\overline{L}_{2(D)}$	.0785	.0764	.0776	.0859	.0864	.0897	.0619	.0504	.0442
					$\pi = .90$				
$\overline{\rho}_{HC}$	.9884	.9877	.9857	.9905	.9902	.9890	.9973	.9981	.9982
$\overline{L}_{2(D)}$	.1420	.1459	.1563	.1309	.1325	.1397	.0703	.0593	.0550
$n = 300$									
					$\pi = .95$				
$\overline{\rho}_{HC}$	.9977	.9979	.9979	.9975	.9976	.9975	.9985	.9991	.9994
$\overline{L}_{2(D)}$	.0648	.0618	.0616	.0674	.0655	.0668	.0495	.0380	.0321
					$\pi = .90$				
$\overline{\rho}_{HC}$	.9929	.9928	.9924	.9905	.9902	.9890	.9985	.9990	.9992
$\overline{L}_{2(D)}$	.1128	.1129	.1157	.1309	.1325	.1397	.0517	.0431	.0378

Table 2: Mean distance and correlations $\overline{L}_{2(D)}$ and $\overline{\rho}_{HC}$. (Simulations I-III).

C.4 Results Simulation Study 2

Since the true structural dimension is $d = 2$, both computational algorithms in Section 2.6 are used, with a slight improvement in the accuracy between the actual and estimated bases using the direct search method. The accuracy results using direct maximization for each level of the tuning parameter $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, with sample size $n = 300$, are reported in Table 3 for brevity.

The mean correlations $\overline{\rho}_{HC}$ in Table 3 are high for both contamination levels, increasing to the highest values as α increases, with the mean $\overline{L}_{2(D)}$ distances decreasing comparably, indicating that our method accurately estimates a more robust basis for $\mathcal{S}_{Y|\mathbf{X}}$ at higher levels of α.

For the successive search algorithm, the AUC_α plots in top right panel of Figure 4 show that the optimal level of the tuning parameter is also in the upper ranges, with $\alpha = 0.8$ producing the most robust index; the plot of the smoothed SIF values in the top left panel confirm this optimal level of the tuning parameter. In addition, the boxplots in the bottom left panel of the |SIF| values $|s_{(i,k)}|$, $k = 1, 2, 3, 4$, clearly estimate the correct structural dimension $\widehat{d} = 2$, with nearly the same center and variability when $k = 2$ as the boxplot when $k = 1$, and a noticeable increased difference compared to $k \geq 3$. The bootstrap procedure of Section 4.2 using the L_2 distance measure in article-(12) is not as obvious, with the boxplots in the bottom right panel showing a more significant increase in both the center and variability from the dimensions $k = 1$ to 2, but together with the notable increase in the centers from $k = 2$ to 3, the estimated structural dimension is again taken to be $\widehat{d} = 2$. Importantly, the improved performance using the actual SIF values comes with less computational effort.

Study 2 Simulation I									
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$n = 300$									
					$\pi = .95$				
$\overline{\rho}_{HC}$	.9712	.9735	.9769	.9784	.9776	.9773	.9779	.9784	.9814
$\overline{L}_{2(D)}$	.1945	.1870	.1798	.1728	.1733	.1716	.1695	.1677	.1599
					$\pi = .90$				
$\overline{\rho}_{HC}$	.9635	.9653	.9660	.9680	.9680	.9702	.9727	.9708	.9747
$\overline{L}_{2(D)}$	.2208	.2149	.2115	.2061	.2012	.1973	.1885	.1908	.1819

Table 3: Mean distance and correlations $\overline{L}_{2(D)}$ and $\overline{\rho}_{HC}$. (Study 2 simulation I; direct search method).

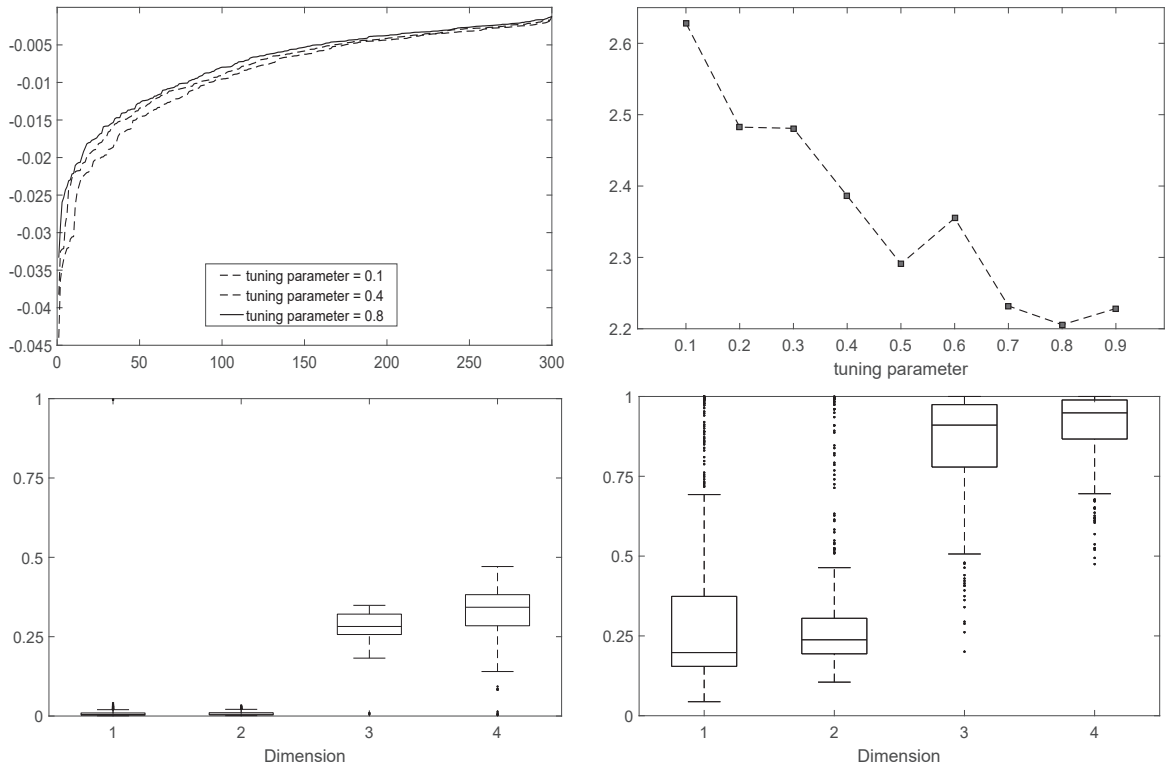

Figure 4: $n = 300, \pi = .90$, $\alpha = 0.8$. Top left panel: smoothed SIF value plots. Top right panel: AUC_α values, $\alpha = 0.1, 0.2, \ldots, 0.9$. Bottom left panel: boxplots of |SIF| values $|s_{(i,k)}|, k = 1, 2, 3, 4$. Bottom right panel: boxplots of bootstrap $L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values, $k = 1, 2, 3, 4$. (Study 2 Simulation I; successive search method).

C.5 Results Simulation Study 3

The accuracy results for each level of the tuning parameter $\alpha \in \{0.1, 0.2, \ldots, 0.9\}$, with sample size $n = 300$, are reported in Table 4 and Table 5 for the direct and successive search methods, respectively.

The mean correlations $\overline{\rho}_{HC}$ in Table 4 are high for both contamination levels and analogously, the $\overline{L}_{2(D)}$ distances are small, indicating that our method accurately estimates a basis for $\mathcal{S}_{Y|\mathbf{X}}$. For both contamination levels the lowest average distance values are consistently observed when $\alpha \geq 0.6$, indicating that these values of the tuning parameter more often parameterized the most robust index.

Next, for a selected dataset with $\pi = 0.90$, the AUC_α plots using the successive search method in the top right panel of Figure 5 show that the optimal α is in the middle ranges, between 0.4 and 0.6, with $\alpha = 0.4$ slightly producing the most robust index. The plot of the smoothed SIF values in the top left panel of Figure 5, and the discussion in Section 1.1 comparing the plots of the SIF values in article-Figure 1 when $\alpha = 0.4$ and $\alpha = .08$, provide further evidence for choosing $\alpha = 0.4$ as the optimal level of the tuning parameter. In addition, the boxplots in the bottom left panel of the |SIF| values $|s_{(i,k)}|$, $k = 1, 2, 3, 4$, clearly estimate the correct structural dimension as $\widehat{d} = 2$, with nearly the same center and variability when $k = 2$ as the boxplot when $k = 1$, and a noticeable increase when compared to $k \geq 3$. The bootstrap procedure of Section 4.2 using the L_2 distance measure in article-(12) is not as obvious, with the boxplots showing a more significant increase in both the center and variability from the dimensions $k = 1$ to 2, but together with the notable increase in the centers from $k = 2$ to 3, the estimated structural dimension is again taken to be $\widehat{d} = 2$. Importantly, this study demonstrates the improved performance using the actual SIF values and consequently, provide a more succinct method for dimension estimation and tuning parameter selection.

Study 3 Simulation I									
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$n = 300$									
				$\underline{\pi = .95}$					
$\overline{\rho}_{HC}$	.9847	.9920	.9872	.9877	.9828	.9911	.9912	.9890	.9947
$\overline{L}_{2(D)}$	.1061	.0984	.1013	.1009	.1057	.0961	.0958	.0974	.0922
				$\underline{\pi = .90}$					
$\overline{\rho}_{HC}$	.9799	.9836	.9915	.9905	.9864	.9918	.9886	.9886	.9846
$\overline{L}_{2(D)}$	.1195	.1144	.1047	.1054	.1085	.1025	.1055	.1052	.1087

Table 4: Mean distance and correlations $\overline{L}_{2(D)}$ and $\overline{\rho}_{HC}$. (Study 3 Simulation I; direct search algorithm).

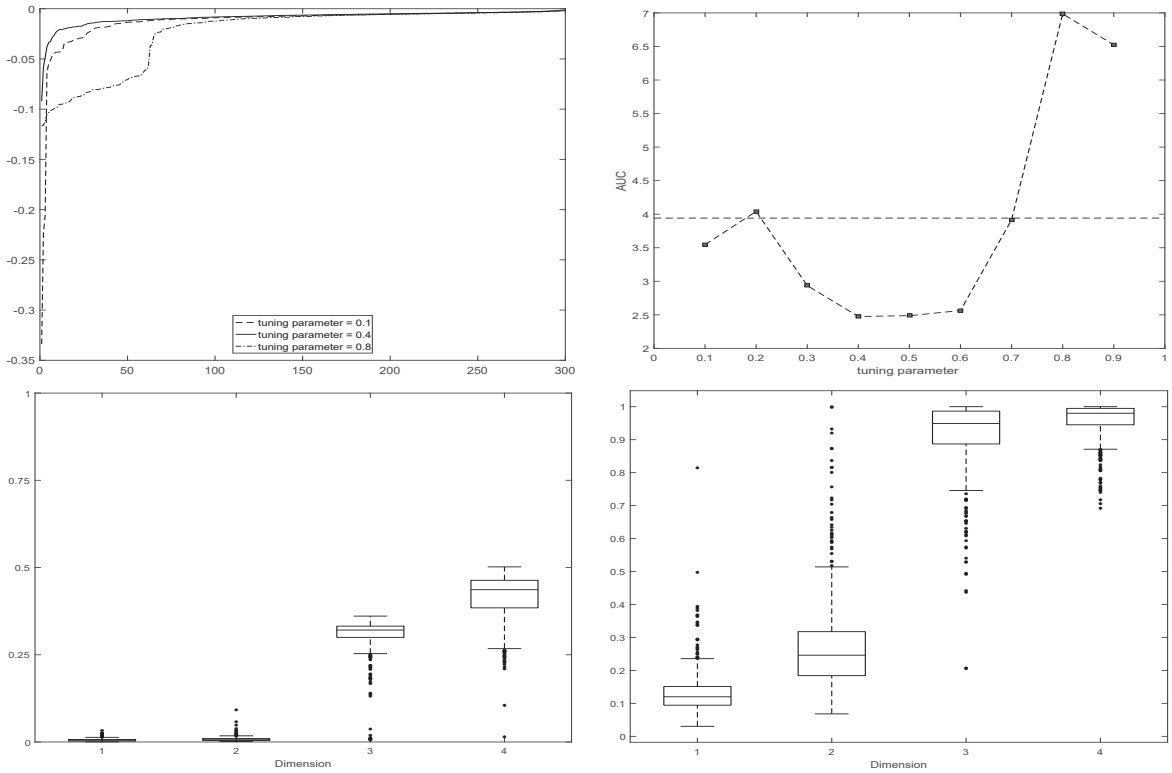

Figure 5: $n = 300, \pi = .90$, $\alpha = 0.4$. Top left panel: smoothed SIF value plots. Top right panel: AUC_α values, $\alpha = 0.1, 0.2, \ldots, 0.9$. Bottom left panel: boxplots of |SIF| values $|s_{(i,k)}|, k = 1, 2, 3, 4$. Bottom right panel: bootstrap boxplots of $L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values, $k = 1, 2, 3, 4$. (Study 3 Simulation I; successive search algorithm).

$\mathcal{R}_\alpha(\mathbf{A})$ **Study 3 Simulation I**									
α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$n = 300$									
				$\underline{\pi = .95}$					
$\overline{\rho}_{HC}$	.9779	.9801	.9808	.9803	.9819	.9797	.9799	.9833	.9834
$\overline{L}_{2(D)}$	.1856	.1791	.1761	.1759	.1704	.1726	.1710	.1630	.1628
				$\underline{\pi = .90}$					
$\overline{\rho}_{HC}$	.9774	.9781	.9762	.9782	.9789	.9786	.9808	.9794	.9809
$\overline{L}_{2(D)}$	.1904	.1879	.1893	.1853	.1820	.1800	.1753	.1760	.1726

Table 5: Mean distance and absolute correlations $\overline{L}_{2(D)}$ and $\overline{\rho}_{HC}$. (Study 3 Simulation I; successive search algorithm).

D Baseball salary data analysis

To illustrate the inherent robustness of our method, we analyze a well-studied dataset that was initially given in a sponsored section on statistics and graphics of the American Statistical Association in 1988, with the stated goal of answering the question; "are players paid according to their performance?" Hoaglin and Velleman (1995) wrote a review of the data analyses performed by the fifteen groups that participated and specifically commented on the considerations taken by authors in dealing with the known outliers and extreme observations present in the dataset. More recently, Xia et al. (2002) analyzed this dataset using their Minimum Average Variance Estimation (MAVE) method for identifying the Effective Dimension Reduction (EDR) subspace in a dimension reduction setting. However, improving the results of their analysis involved first identifying outliers, and then removing the observations deemed influential. Different from the previously mentioned analyses for predicting annual salary from the predictors, our procedure to estimate regression DR directions does not require a preliminary analysis to identify outliers, which is inherently difficult in high dimensional settings. In addition, the directions can be used to create linear combinations of the predictors to build predictive models for further analysis. Note that the following analysis is performed in the whitened scale, but the notations $\mathbf{X}$ and Y are maintained for continuity.

The random vector for predicting annual salary, $\mathbf{X} = (X_1, X_2, \ldots, X_{16})^\top$, consists of the variables: times at bat X_1, hits X_2, home runs X_3, runs X_4, runs batted in X_5, walks X_6, errors X_7, putouts X_8, and assists X_9, in the 1986 season. The remaining *career* predictor variables are the number of: times at bat X_{10}, hits X_{11}, home runs X_{12}, runs X_{13}, runs batted in X_{14}, walks X_{15}, and years in the major leagues X_{16}, for the players career up to the 1986 season. The dependent variable Y is the annual salary in 1986 in natural log scale.

The boxplots of the |SIF| values $|s_{(i,k)}|$, $k = 1, 2, 3, 4$, for implementing the dimension detection method described in Section 4.3 are displayed in the right panel of Figure 7, and clearly identify the existence of one strong relationship between the response and explanatory vector, which is analogously supported by the boxplots of the bootstrapped $1 - \rho_{HC}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$ values detailed in Section 4.2 in the left panel of the same figure. Arguably, both methods indicate that a meaningful second regression direction exists, although weaker than the first direction, as demonstrated by the boxplots having slightly higher means with more variability in comparison. The difference in the strengths of these relationship is apparent in the smoothed scaled SIF value plots in the left and right panels of Figure 8 for $k = 1$ and 2, respectively. However, taken together with the nonlinear relationship visible in the variate plot of the observed response y versus the variate $\widehat{\mathbf{a}}_2^\top \mathbf{x}$ in the middle panel of Figure 6, the estimated dimension of $\mathcal{S}_{Y|\mathbf{X}}$ is taken to be $\widehat{d} = 2$; using the second regression direction further ensures no loss of information between the response and predictors. The step function nature of the SIF value plots in Figure 8 for $k = 1$ indicates that about 50 of the $n = 263$ observations are the most influential.

Consider the loadings for the 1st coefficient vector, or regression dimension reduction direction, $\widehat{\mathbf{a}}_1$ given in Table 6. The variables hits in the 1986 season X_2 and the career number of times at bat X_{10} are given the most significant positive weights of .761 and .958, respectively. The next largest, but significantly less, weight of .101 is placed on the predictor X_4 runs in the 1986 season, which is naturally positively associated with X_2 and X_{10}. The remaining variables are given even less weight and thus, conclude that the 1st estimated loadings of the predictor vector $\mathbf{X}$ are overwhelmingly a weighted average of X_2 and X_{10}. The plot of y versus $\widehat{\mathbf{a}}_1^\top \mathbf{x}$ in the left panel of Figure 6 indicates that a strong nonlinear relationship with a quadratic trend is recovered. Specifically, the annual salary is lowest when the variate is less than -1 and then increases sharply in a tight linear manner

as the variate approaches 0, which likely corresponds to players progressing from the start to the middle of their careers when their total number of times at bat are of more moderate values. This linear trend remains for variate values between 0 and 1, with much more variation, and then trends downward as the variate increases past 2, creating an overall quadratic trend for variate values exceeding 1. The higher values between 0 and 2 are likely players who have been in the league for a longer time, which seems reasonable as these players would be expected to have larger observed career number of times at bat; anticipated with the increased number of times at bat are larger differences in the number of hits in the 1986 season, which would explain the increased variability. The downward trend for variate values greater than 2 is likely attributed to players at the end of their careers, which has been previously explained as an aging effect; likely reflected here in a more significant reduction in hitting in the 1986 season and consequently, fewer runs.

For the 2$^{\text{nd}}$ coefficient vector, the loadings in table 6 show that the largest weights, from -.594. to -.524, are given to the variables X_{11}, X_{13} and X_{14}; the career totals for number of hits, runs, and runs batted in. Note that, the career homeruns X_{12} is strongly associated with theses variables, which is reflected in the lower weight of -.185. There is arguably a weak contrast between these predictors and X_{16} number of years in the league, X_2 hits and X_4 runs in the 1986 season, with weights .125, .154 and .208, respectively. However, the coefficients predominately load on career batting statistics, and since career runs and runs batted in are a function of the career number of hits, this can be viewed as a career *batting* coefficient vector. The variates plots in the middle panel of Figure 6 again reveal an overall nonlinear relationship between the batting variates $\widehat{\mathbf{a}}_2^\top \mathbf{x}$ and the annual salary. When the value of the variate is between -.5 to 1 there is a strong increasing linear relationship with annual salaries less than 0, which likely corresponds to players in the beginning of their careers and thus, lower career batting totals. This increasing linear trend in annual salaries continues for variate values greater than 0, which can be generally attributed to players in the beginning of their career who have lower total career batting totals but high season batting statistics and thus, players who are expected to have high career batting totals by the end of their careers. The highest salaries are generally observed for variate values less than -1, which creates the observed quadratic trend, and corresponds to players with higher career batting totals.

The 1$^{\text{st}}$ variate plot in Figure 6 displays a similar nonlinear trend to the analogous graph based on the estimated EDR direction in Figure 6 of Xia et al. (2002), after the removal of seven observations that were deemed outliers. However, there are some distinct differences between the loadings of the regression DR directions. The greatest positive weights for the EDR direction are placed on the career variables, number of years in the majors, times at bat, hits, and walks, with values .52, .55, .37 and .30, respectively, However, the Rényi divergence based direction puts the largest weight of .958 on the career number of times at bat X_{10}, which can viewed as a summary of the EDR weights, since, for example, players with an observed large number of at bats would also be expected to have a higher amount of hits, more at bats, more walks and been in the major leagues longer. Interestingly, the next EDR weights in this coefficient vector provide a contrast between the number of times at bat and hits in the 1986 season, with values -.25 and .24, where as, the Rényi divergence based directions puts the last significant large weight of .761 on the number of hits in 1986 X_2. Therefore, using our method the 1$^{\text{st}}$ calculated direction provides a more parsimonious and thus, more interpretable coefficient vector without the need for the identification and exclusion of extreme observations. As in Xia et al. (2002), after determining the two variates, $\widehat{\mathbf{a}}_1^\top \mathbf{x}$ and $\widehat{\mathbf{a}}_2^\top \mathbf{x}$, we fit a linear model using the two variates as predictors with stepwise linear regression producing the fitted model $\widehat{y} = 0.42672 + 0.96824(\widehat{\mathbf{a}}_1^\top \mathbf{x}) - 0.228(\widehat{\mathbf{a}}_2^\top \mathbf{x}) - .42835(\widehat{\mathbf{a}}_1^\top \mathbf{x})^2$. Note that, Xia et al. (2002) also reported an r^2 value of 0.714 for their model fitted using the EDR directions. In comparison, the adjusted r^2 for our model is 0.767. Therefore, our method is shown to effectively

mitigate the effect of the well established outlying observations present in this dataset without their identification and removal.

Hitter Data Analysis – $\mathcal{R}_{0.1}(\mathbf{A})$																
	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}
$\widehat{\mathbf{a}}_1$	.041	**.761**	.008	.101	.046	.090	-.046	.035	-.006	**.958**	.032	.064	-.019	.048	.095	.091
$\widehat{\mathbf{a}}_2$	-.093	.125	-.010	.154	.046	-.021	-.021	.005	.001	-.013	**-.575**	-.185	**-.424**	**-.594**	.095	.208

Table 6: Table of estimated coefficient vector loadings (Example Baseball salary).

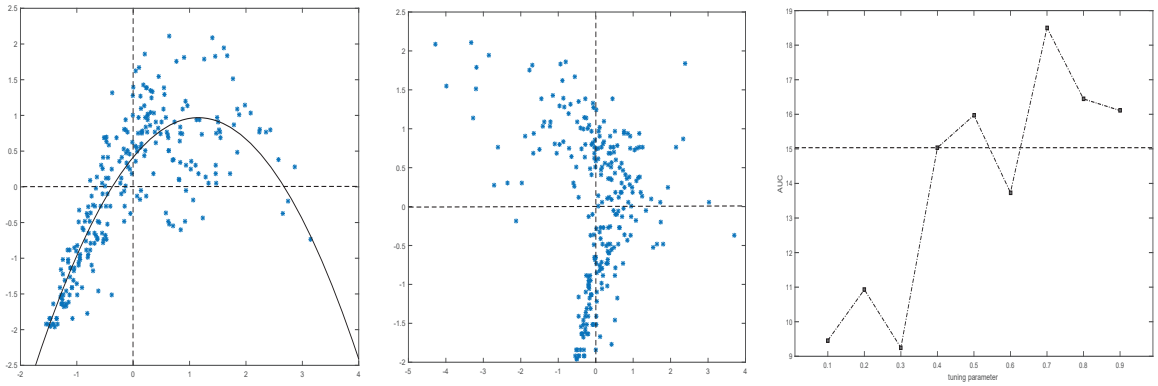

Figure 6: Right panel: $\widehat{\mathbf{a}}_1^\top \mathbf{x}$ vs. y, $\alpha = 0.1$. Middle panel: $\widehat{\mathbf{a}}_2^\top \mathbf{x}$ vs. y, $\alpha = 0.1$. Right panel: AUC_α values, dimension $k = 1$, $\alpha = 0.1, 0.2, \ldots, 0.9$

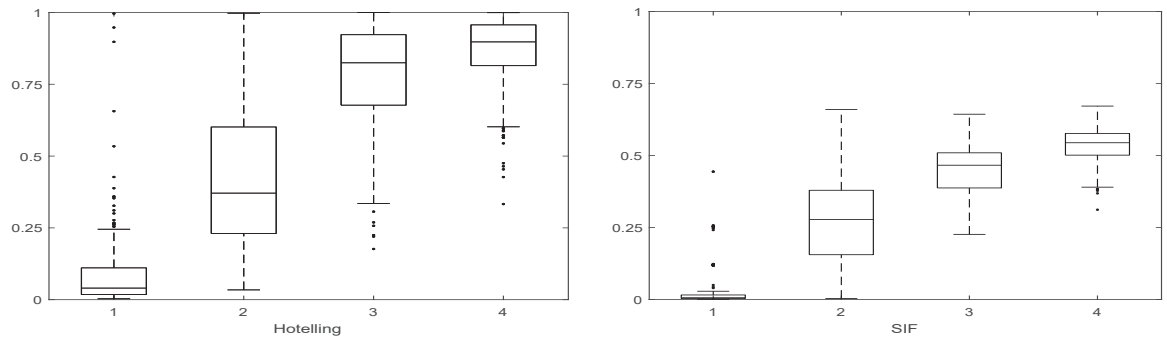

Figure 7: Boxplots, $\alpha = 0.1$, dimension $k = 1, 2, 3, 4$. Left Panel: Bootstrap $1 - \rho_{HC}\left(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\right)$ values. Right panel: |SIF| values $|s_{(i,k)}|$.

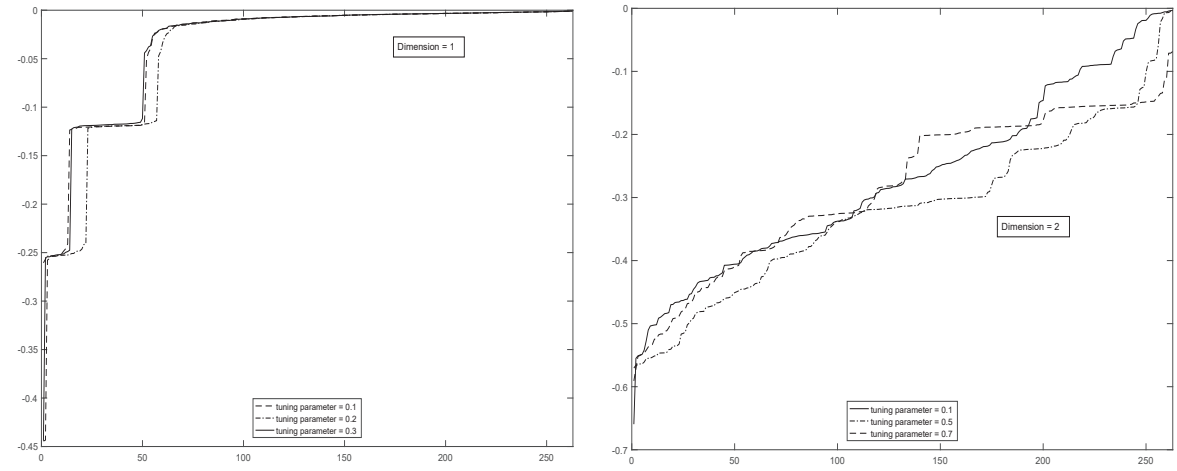

Figure 8: Smoothed SIF value plots. Left panel: dimension $k = 1$, $\alpha = 0.1, 0.2, 0.3$. Right panel: dimension $k = 2$, $\alpha = 0.1, 0.5, 0.7$. (Baseball salary data analysis)

E Additional Simulation Study 4

E.1 Introduction

For this comparative study, we consider the same distribution and covariance structure of the predictors, error term, symmetric outliers generated from a uniform distribution, and follow their simulation parameters, including using their measure for quantifying the accuracy of the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$, for the regression model in Study 3. The changes in the simulation design are detailed next.

The accuracy between the true and estimated subspaces is measured using the trace correlation coefficient of Hooper (1959), defined as

$$\rho_{TR}\big(\widehat{\mathbf{A}},\mathbf{A}\big)=\sqrt{\text{trace}\big\{(\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^T\widehat{\mathbf{A}}(\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}})^{-1}\widehat{\mathbf{A}}^\top\mathbf{A}\big\}/d}=\sqrt{\text{trace}\big\{\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}\big\}/d}=\left(\frac{1}{d}\sum_{i=1}^{\text{d}}\lambda_i\right)^{\frac{1}{2}}. \tag{E.1}$$

As in Section 4.2, the λ_i are the eigenvalues of $\mathbf{A}^\top\widehat{\mathbf{A}}\widehat{\mathbf{A}}^\top\mathbf{A}$, $0\leq\rho_{TR}\big(\widehat{\mathbf{A}},\mathbf{A}\big)\leq 1$, $\rho_{TR}\big(\widehat{\mathbf{A}},\mathbf{A}\big)=1$ implies that $\mathcal{S}(\mathbf{A})$=$\mathcal{S}(\widehat{\mathbf{A}})$, and $\rho_{HC}\big(\widehat{\mathbf{A}},\mathbf{A}\big)=0$ when the subspaces are orthogonal. Note that, $0\leq\lambda_i\leq 1$ (Ye and Weiss (2003)), which for $d>1$ implies that $\rho_{TR}\big(\widehat{\mathbf{A}},\mathbf{A}\big)\geq\rho_{HC}\big(\widehat{\mathbf{A}},\mathbf{A}\big)$ since $\prod_{i=1}^d\lambda_i\leq\prod_{i=1}^d\lambda_i^{\frac{1}{d}}=\big(\prod_{i=1}^d\lambda_i\big)^{\frac{1}{d}}\leq\frac{1}{d}\sum_{i=1}^d\lambda_i$, where the last inequality follows from the arithmetic-geometric mean inequality.

Uncontaminated error terms in the regression model are generated from a $N(0,1)$ distribution, and symmetric outlying observations produced at random from a $U(-\theta,\theta)$ distribution with probability $\pi=.95$. Importantly, Zhang, Wang and Mays (2021) defined the error term in the regression model as $0.5\varepsilon^*$, where $\varepsilon^*\sim U(-\theta,\theta)$, which would be equivalent to generating a contaminated error term from a $U\big(-0.5\theta,0.5\theta\big)$ distribution in our previous simulation setup. As in Sections C.2, this is denoted $\varepsilon^*\sim N(0,1)\mathcal{I}(\pi)+U(-\theta,\theta)\{1-\mathcal{I}(\pi)\}$, where $\mathcal{I}(\pi)=1$ with probability π.

Different from Simulation Study 3 in Section C.5, where the predictor vector is composed almost entirely of variables that follow a variety of skewed and heavy-tailed distributions, the explanatory vector is multivariate normal with a Toeplitz matrix covariance dependence structure. The distribution of the predictor vector and model error terms are summarized as:

> Study 4: $\mathbf{X}=(X_1,X_2,\ldots,X_{10})^\top\sim N_{10}(\mathbf{0},\Sigma)$, where the (i,j)th term of Σ is $\sigma_{ij}=0.5^{|i-j|}$; $\varepsilon^*\sim N(0,1)\mathcal{I}(\pi)+U(-\theta,\theta)\{1-\mathcal{I}(\pi)\}$, $\pi=.95$.

The regression model with the newly defined error term and true coefficient matrix are given in Table 7.

Study 4	Model	True Coefficient Matrices
	$Y=\frac{\mathbf{a}_1^\top\mathbf{X}}{0.5+\left(\mathbf{a}_2^\top\mathbf{X}+1.5\right)^2}+0.5\varepsilon^*$	$\mathbf{A}=[(1,0,\ldots,0)^\top;(0,1,0,\ldots,0)^\top]$

Table 7: Simulation regression model simulation Study 4.

Two simulation studies are performed to investigate the effect of symmetric contamination on the Rényi based method. Simulation I is the comparative simulation with Zhang, Wang and Mays (2021), where $\varepsilon^* \sim U(-50, 50)$. In Simulation II, we increase the magnitude of contamination in Simulation II by setting $\theta = 100$ and thus, $\varepsilon^* \sim U(-100, 100)$.

For a dataset generated under the above specifications, $\widehat{\mathbf{A}}$ is calculated and the difference between the true coefficient matrix $\mathbf{A}$ is quantified using (E.1). For datasets of size $n = 100, 200, 300$ and 400, this process is repeated $n_s = 200$ times and the overall accuracy in estimating $\mathbf{A}$ is measured by $\overline{\rho}_{TR} = \frac{1}{n_s}\sum_{j=1}^{n_s} \rho_{HC}(\widehat{\mathbf{A}}^j, \mathbf{A})$, where $\widehat{\mathbf{A}}^j$ is the estimated coefficient matrix for the j^{th} simulated dataset; note that, Zhang, Wang and Mays (2021) did not include $n = 300$ in their simulation study.

To visualize the effect of the symmetric contamination in comparison to the previously investigated asymmetric outliers, a randomly selected simulated dataset of size $n = 300$ is plotted in the left panel of Figure 9, with the asymmetric contamination plot of Section C.1 Study 3 reproduced in the right panel. Note that, in terms of the currently defined error term, this would be equivalent to $\varepsilon^* \sim U(0, 40)$, and that the magnitude of the effect of the asymmetric outliers on the response in the positive direction is comparable. In the direct comparison numerical study to that Zhang, Wang and Mays (2021), Simulation I, the Rényi divergence based method has a discernible higher mean trace correlation for all sample sizes in the presence of asymmetric contamination; see Section E.2. However, the effect of asymmetric contamination, as shown in the right panel of Figure 9, was not addressed in their paper. Note that, for both simulations the standard errors of the means are less than 10^{-2} and not reported in the tables of the results in Section E.2 for brevity .

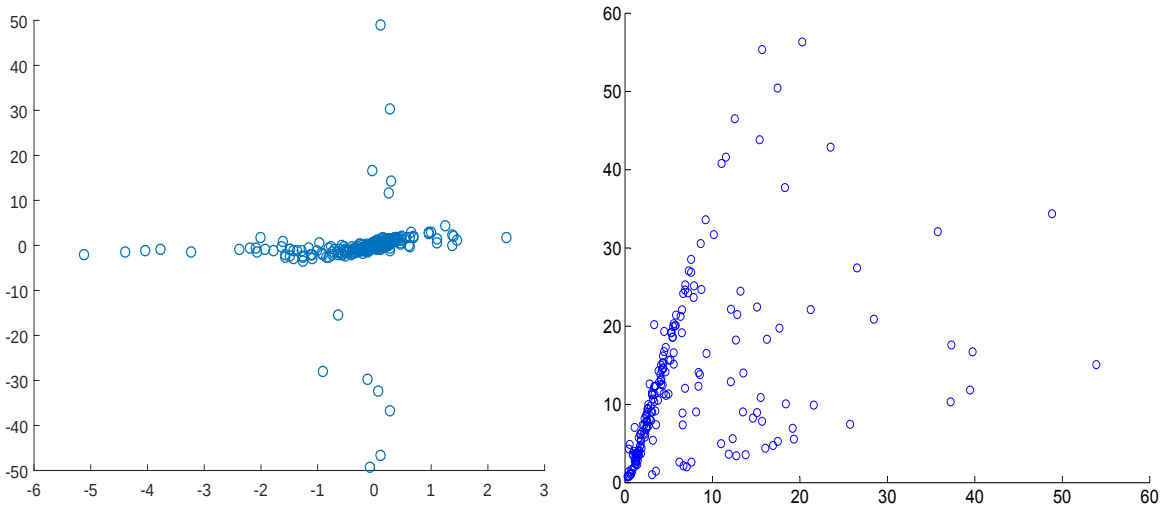

Figure 9: Symmetric vs. Asymmetric Outliers Study 3 and Study 4 Simulation II $\{\varepsilon^* \sim U(-100, 100)\}$, $\pi = .95$, example data plot y versus $\mathbf{a}_1^\top \mathbf{x}/\{0.5 + (\mathbf{a}_2^\top \mathbf{x} + 1.5)^2\}$. Left panel: Study 4 (symmetric) $n = 300$. Right panel: Study 3 (asymmetric) $n = 200$.

E.2 Results Simulations I and II

The mean trace correlations $\overline{\rho}_{TR}$ for Simulations I and II at each sample size are reported in Table 8. For both simulations the most robust levels of the tuning parameter are attained on average for smaller values of tuning parameter, between 0.1 and 0.5, with a more noticeable separation in Simulation II when the contaminated error terms are generated from a $U(-100, 100)$. In general,

the values of $\alpha = 0.1$ and 0.2 provide the highest mean trace correlations. In contrast, the most robust levels of α in the asymmetric outlier simulation in study 3 were typically from $\alpha = .4$ to $\alpha = .6$, which demonstrates that the range of the tuning parameters are capable of handling both symmetric and asymmetric outliers.

In Simulation I with $\varepsilon^* \sim U(-50, 50)$, the mean trace correlation values are notably higher using the Rényi divergence based method, with values .8646, .9849 and .9904, compared to the best results reported in Table 1 in Zhang, Wang and Mays (2021) corresponding to their L_0 penalty based method, with values .7869, .9055 and .9685, at the comparable sample sizes $n = 100, 200$ and 400, respectively. Moreover, the Rényi based mean trace correlations, when $\alpha = 0.1$ for example, remain comparable to the values of Zhang, Wang and Mays (2021) ($\varepsilon^* \sim U(-50, 50)$) when the contamination is doubled to $\varepsilon^* \sim U(-100, 100)$ in Simulation II. Importantly, this simulation study demonstrates the ability of our method to reliably estimate $\mathcal{S}_{Y|\mathbf{X}}$ in the presence of both collinearity and symmetric outliers.

Study 4 **Mean Trace Correlations** $\overline{\rho}_{_{TR}}$

Simulation I: $\varepsilon^* \sim U(-50, 50)$, $\pi = .95$

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
n									
100	.8646	.8613	.8549	.8561	.8515	.8441	.8446	.8291	.8204
200	.9662	.9658	.9639	.9630	.9594	.9494	.9446	.9337	.9202
300	.9849	.9850	.9850	.9845	.9842	.9836	.9808	.9790	.9740
400	.9904	.9906	.9906	.9907	.9905	.9902	.9890	.9880	.9869

Simulation II: $\varepsilon^* \sim U(-100, 100)$, $\pi = .95$

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
n									
100	.7782	.7687	.7719	.7651	.7621	.7501	.7442	.7340	.7195
200	.8851	.8826	.8798	.8708	.8593	.8452	.8331	.8126	.7893
300	.9416	.9400	.9336	.9272	.9194	.9064	.8895	.8676	.8411
400	.9684	.9659	.9648	.9628	.9550	.9462	.9371	.9189	.9021

Table 8: Mean trace correlations $\overline{\rho}_{_{TR}}$. (Study 4 simulations I and II; direct search method).

References

[1] Hoaglin, D. and Velleman, P. (1995). A critical look at some analyses of major league baseball salaries, The American Statistician 49, 277-285.

[2] Hotelling, H. (1936). Relations between two sets of variables, Biometrika 58, 433–51.

[3] Hooper, J. (1959). Simultaneous Equations and Canonical Correlation Theory, Econometrica 27, 245-256.

[4] Prendergast, L. A. (2006). Detecting influential observations in Sliced Inverse Regression analysis, Aust. N. Z. J. Statist. 48, 285–304.

[5] Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002). An adaptive estimation of dimension reduction, J. R. Statist. Soc. B 64, 363–410.

[6] Ye, Z. and Weiss, R.E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods, J. Amer. Statist. Assoc. 98, 363–410.

[7] Yin, X. and Cook, R.D. (2005). Direction Estimation in Single-Index Regressions, Biometrika 92, 371–384.

[8] Zhang, J., Wang, Q. and Mays, D. (2021). Robust MAVE through nonconvex penalized regression, Comp. Statist. & Data Analysis 160, 107247.