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

T. N. Sriram

Department of Statistics

University of Georgia

Co-author: Ross Iaci

We dedicate this work to our late colleague Professor Xiangrong Yin.

June 26, 2023

Notations

- Y is a univariate response variable
- $\boldsymbol{X}$ is a $p \times 1$ vector of predictors
- $(Y_i, \boldsymbol{X}_i)$, $i = 1, 2, \ldots, n$, are i.i.d. observations of $(Y, \boldsymbol{X})$
- $(Y, \boldsymbol{X}) \sim f(y, \boldsymbol{x})$ — ioint density function
- $F(y, \boldsymbol{x})$ — the joint distribution function

Goal

Goal of dimension reduction:

- Find k coefficient vectors (or directions) $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k$ such that significant relationships between Y and $\boldsymbol{X}$ are identified through the k linear combinations $\boldsymbol{a}_1^\top \boldsymbol{X}, \ldots, \boldsymbol{a}_k^\top \boldsymbol{X}$, where $1 \leq k < p$.

Dimension Reduction Subspace

- Let $\mathcal{S}(\boldsymbol{B})$ be the k-dimensional subspace in $\boldsymbol{R}^p$ spanned by the columns the matrix $\boldsymbol{B}$
- $P_{\mathcal{S}(\boldsymbol{B})}$ the projection onto $\mathcal{S}(\boldsymbol{B})$
- $\mathcal{S}(\boldsymbol{B})$ is a Dimension Reduction Subspace (DRS) for the regression of Y on $\boldsymbol{X}$ if

$$Y \perp\!\!\!\perp \boldsymbol{X} | P_{\mathcal{S}(\boldsymbol{B})} \boldsymbol{X} \text{ [Conditional Independence]}$$

- That is, $\mathcal{S}(\boldsymbol{B})$ is a DRS if $f(y|\ \boldsymbol{x}) = f(y|\ P_{\mathcal{S}(\boldsymbol{B})}\boldsymbol{x})$, for all $(y, \boldsymbol{x}) \in \mathrm{R} \times \mathrm{R}^p$.

Central Subspace (CS)

- Conditional independence holds if $\boldsymbol{B}$ is replaced with any matrix $\boldsymbol{B}^*$ such that $\mathcal{S}(\boldsymbol{B}^*) = \mathcal{S}(\boldsymbol{B})$, which means that any basis of a DRS is also a DRS.
- When $\mathcal{S}(\boldsymbol{B})$ is a DRS, the transformation $\boldsymbol{B}^\top \boldsymbol{X}$ provides a *sufficient* dimension reduction
- Let $\mathcal{S}_{Y|\boldsymbol{X}}$ denote the intersection of all DRSs, which is a DRS under mild conditions
- $\mathcal{S}_{Y|\boldsymbol{X}}$ is called the Central Subspace (CS)

Basis of CS

- Let $d = dim(\mathcal{S}_{Y|\boldsymbol{X}}) < p$ denote the true dimension of the CS.
- d is called the structural dimension of Y on $\boldsymbol{X}$.
- Assume, $\boldsymbol{A}$ is a d-dimensional basis for $\mathcal{S}_{Y|\boldsymbol{X}}$
- Then the conditional distributions of $Y|\boldsymbol{A}^\top \boldsymbol{X}$ and $Y|\boldsymbol{X}$ are the same
- We assume that $\mathcal{S}_{Y|\boldsymbol{X}}$ exists with structural dimension d and focus on the **robust estimation** of a basis $\boldsymbol{A}$ for $\mathcal{S}_{Y|\boldsymbol{X}}$.

Literature Review

- Pioneering methods: Sliced Inverse Regression (SIR) and Sliced Average Variance Estimation (SAVE).
- There are many powerful model-free dimension reduction methods in the literature.
- While useful, many DR methods are highly sensitive to influential observations.
- There are studies on the sensitivity of the existing DR methods to extreme observations.
- This led to construction of robust version of existing methods.
- Minumum Average Variance Estimation (MAVE) method is also not robust against outliers in the response variable Y.

Outliers

- Lack of robustness of DR methods are exacerbated for high dimensional (HD) datasets
- It is not only difficult to detect outlying and/or influential observations in HD but often hard to resolve when they are identified.
- There are studies providing ways of assessing the influence of extreme observations on the estimates provided by SIR, pHd and SAVE
- But they do not provide a way to construct estimates that are inherently robust to data contamination.

Rényi Divergence

- **Our goal** is to provide a comprehensive methodology, based on the Rényi divergence that is inherently robust to data contamination.
- For $\alpha > 0$, $\alpha \neq 1$, $f_1(\boldsymbol{u})$ & $f_2(\boldsymbol{u})$ p.d.f.s, Rényi divergence is:

$$D_\alpha\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\} = \frac{1}{\alpha - 1} \ln\left[\int_{\boldsymbol{u}} \left\{\frac{f_1(\boldsymbol{u})}{f_2(\boldsymbol{u})}\right\}^{\alpha-1} f_1(\boldsymbol{u})\, d\boldsymbol{u}\right].$$

- $D_\alpha\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\} \geq 0$
- $D_\alpha\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\} = 0$ if and only if $f_1(\boldsymbol{u}) = f_2(\boldsymbol{u})$.

Two special cases of Rényi Divergence

Kullback Leibler (KL) divergence is a limiting case:

$$\lim_{\alpha \to 1} D_\alpha\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\} = \int_{\boldsymbol{u}} \ln\left\{\frac{f_1(\boldsymbol{u})}{f_2(\boldsymbol{u})}\right\} f_1(\boldsymbol{u})\, d\boldsymbol{u} = D_{KL}\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\}$$

- For $\alpha = 1/2$, $D_{1/2}\{f_1(\boldsymbol{U})||f_2(\boldsymbol{U})\} = -2\ln\left[1 - \left\{(HB)^2/2\right\}\right]$,
- $HB = \left[\int_{\boldsymbol{u}} \left\{\sqrt{f_1(\boldsymbol{u})} - \sqrt{f_2(\boldsymbol{u})}\right\}^2 d\boldsymbol{u}\right]^{1/2}$ is the Hellinger-Bhattacharyya (HB) distance.

Robust Identification of the CS

For each $\alpha \in (0,1)$, consider a new Rényi divergence-based index

$$\mathcal{R}_\alpha(\boldsymbol{A}) = D_\alpha\{f(Y, \boldsymbol{A}^\top \boldsymbol{X}) || f(Y) f(\boldsymbol{A}^\top \boldsymbol{X})\}$$
$$= \frac{1}{\alpha - 1} \ln \left[\int_y \int_{\boldsymbol{A}^\top \boldsymbol{x}} \left\{ \frac{f(y, \boldsymbol{A}^\top \boldsymbol{x})}{f(y) f(\boldsymbol{A}^\top \boldsymbol{x})} \right\}^{\alpha - 1} f(y, \boldsymbol{A}^\top \boldsymbol{x}) \, d(\boldsymbol{A}^\top \boldsymbol{x}) \, dy \right].$$

Note that

$$\mathcal{R}_\alpha(\boldsymbol{A}) \geq 0.$$

Also,

$$\mathcal{R}_\alpha(\boldsymbol{A}) = 0 \iff f(y, \boldsymbol{A}^\top \boldsymbol{x}) = f(y) f(\boldsymbol{A}^\top \boldsymbol{x}), \text{ i.e., } Y \perp\!\!\!\perp \boldsymbol{A}^\top \boldsymbol{X}.$$

Key Properties of $\mathcal{R}_\alpha(\boldsymbol{A})$

Proposition 1: Let $\boldsymbol{A}$ and $\boldsymbol{A}_1$ denote $p \times k$ and $p \times l$ matrices, with $k, l \leq p$. For $\alpha \in (0,1)$, and a $p \times p$ identity matrix $\mathcal{I}$, then the following hold:

(i) If $\mathcal{S}(\boldsymbol{A}_1) \subseteq \mathcal{S}(\boldsymbol{A})$, *then* $\mathcal{R}_\alpha(\boldsymbol{A}_1) \leq \mathcal{R}_\alpha(\boldsymbol{A})$.

(ii) If $\mathcal{S}(\boldsymbol{A}_1) = \mathcal{S}(\boldsymbol{A})$, *then* $\mathcal{R}_\alpha(\boldsymbol{A}_1) = \mathcal{R}_\alpha(\boldsymbol{A})$.

(iii) $\mathcal{R}_\alpha(\boldsymbol{A}) \leq \mathcal{R}_\alpha(\mathcal{I})$, *and*

(iv) $\mathcal{R}_\alpha(\mathcal{I}) = \mathcal{R}_\alpha(\boldsymbol{A})$ *if and only if* $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{A}^\top \boldsymbol{X}$.

Implications of Proposition 1

- $\mathcal{S}(\boldsymbol{A}_1) \subseteq \mathcal{S}(\boldsymbol{A})$, *then* $\mathcal{R}_\alpha(\boldsymbol{A}_1) \leq \mathcal{R}_\alpha(\boldsymbol{A})$ says searches made successively through increasing dimensional subspaces will ultimately yield a basis for $\mathcal{S}_{Y|\boldsymbol{X}}$ when $Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{A}^\top \boldsymbol{X}$ is satisfied.
- $\mathcal{S}(\boldsymbol{A}_1) = \mathcal{S}(\boldsymbol{A})$, *then* $\mathcal{R}_\alpha(\boldsymbol{A}_1) = \mathcal{R}_\alpha(\boldsymbol{A})$ says matrices that span the same subspace have the same measured dependence and therefore, only a basis for the subspace is needed.
- The bound $\mathcal{R}_\alpha(\boldsymbol{A}) \leq \mathcal{R}_\alpha(\mathcal{I})$ indicates that the largest dependence between $\boldsymbol{X}$ and Y in k dimensions can be recovered by maximizing $\mathcal{R}_\alpha(\boldsymbol{A})$ with respect to $\boldsymbol{A}$.
- Finally, if $\mathcal{R}_\alpha(\boldsymbol{A}) = \mathcal{R}_\alpha(\mathcal{I})$ for a fixed α, then $\boldsymbol{A}$ provides a basis for a k-dimensional DRS in $\boldsymbol{R}^p$ and accordingly, $\boldsymbol{A}^\top \boldsymbol{X}$ is a sufficient dimension reduction for the regression of Y on $\boldsymbol{X}$.

What is A?

- When $\mathcal{R}_\alpha(\boldsymbol{A}) = \mathcal{R}_\alpha(\mathcal{I})$, $\boldsymbol{A}^\top \boldsymbol{X}$ provides a minimum sufficient dimension reduction. That is, for d known and α fixed, a basis $\boldsymbol{A}_{p\times d}$ for $\mathcal{S}_{Y|\boldsymbol{X}}$ can be recovered as

$$\boldsymbol{A} = \arg\max \mathcal{R}_\alpha(\boldsymbol{A}^*) \quad \text{subject to the constraint} \quad \boldsymbol{A}^\top \Sigma_{\boldsymbol{X}} \boldsymbol{A} = \mathcal{I},$$

  where $\Sigma_{\boldsymbol{X}}$ is the covariance matrix of the explanatory vector $\boldsymbol{X}$.

- **We will estimate A without assuming a model.**

Sample estimation

Consider a random sample $\{(y_i, \boldsymbol{x}_i); i = 1, 2, \ldots, n\}$ from $(Y, \boldsymbol{X})$, and assume that the structural dimension d of the regression is known. for any $p \times k$ matrix $\boldsymbol{A}$ and $\alpha \in (0, 1)$, we define the following sample estimate

$$\widehat{\mathcal{R}}_\alpha(\boldsymbol{A}) = \frac{1}{\alpha - 1} \ln \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\widehat{f}(y_i, \boldsymbol{A}^\top \boldsymbol{x}_i)}{\widehat{f}(y_i)\widehat{f}(\boldsymbol{A}^\top \boldsymbol{x}_i)} \right\}^{\alpha - 1} \right],$$

where $\widehat{f}(y_i)$, $\widehat{f}(\boldsymbol{A}^\top \boldsymbol{x}_i)$ and $\widehat{f}(y_i, \boldsymbol{A}^\top \boldsymbol{x}_i)$ are kernel density estimates of $f(y_i)$, $f(\boldsymbol{A}^\top \boldsymbol{x}_i)$ and $f(y_i, \boldsymbol{A}^\top \boldsymbol{x}_i)$, respectively.

Sample estimation

Specifically, to estimate $f(y_i, \boldsymbol{A}^\top \boldsymbol{x}_i)$ for a specific coefficient matrix $\boldsymbol{A} = [\boldsymbol{a}_1\ \mathrm{a}_2 \cdots \boldsymbol{a}_k]$, we use the Gaussian product kernel density estimate

$$\widehat{f}(y_i, \boldsymbol{A}^\top \boldsymbol{x}_i) = \frac{1}{nh^* \prod_{l=1}^{k} h_l} \sum_{j=1}^{n} \left(K\left[\{(y_j - y_i)\}/h^*\right] \prod_{l=1}^{k} K\left[\{\boldsymbol{a}_l^\top (\boldsymbol{x}_j - \boldsymbol{x}_i)\}/h_l\right] \right),$$

with bandwidths $h^* = (4/3)^{1/5}\, s_y\, n^{-1/5}$ and

$h_l = \{4/(k+2)\}^{1/(k+4)}\, s_l\ n^{-1/(k+4)}$, $l = 1, 2, \ldots, k$, where s_y and s_l are the sample standard deviations of the sample observations $\{y_i, i = 1, \ldots, n\}$ and $\{\boldsymbol{a}_l^\top \boldsymbol{x}_i, i = 1, \ldots, n\}$, respectively.

Estimator

For $\alpha \in (0,1)$, our Rényi divergence based estimator of $\boldsymbol{A}$ is defined as

$$\widehat{\boldsymbol{A}} = \operatorname{argmax} \widehat{\mathcal{R}}_\alpha(\boldsymbol{A}^*) \quad \text{subject to the constraint} \quad \widehat{\boldsymbol{A}}^\top \widehat{\Sigma}_{\boldsymbol{x}} \widehat{\boldsymbol{A}} = \mathcal{I},$$

where $\widehat{\Sigma}_{\boldsymbol{X}}$ is the sample estimate of the covariance matrix of $\boldsymbol{X}$.

Main Theorem

Theorem (Consistency)

Let $\widehat{\boldsymbol{A}} = \arg\max \widehat{\mathcal{R}}_\alpha(\boldsymbol{A}^*)$ *and* $\boldsymbol{A} = \arg\max \mathcal{R}_\alpha(\boldsymbol{A}^*)$, *for each* $\alpha \in (0, 1)$,

$$\widehat{\boldsymbol{A}} \to \boldsymbol{A} \text{ as } n \to \infty \text{ almost surely (a.s.)}.$$

Heuristic argument for robustness

When α is close to 1, by L'Hospital's rule

$$\widehat{\mathcal{R}}_\alpha(\boldsymbol{A}) = \frac{1}{\alpha-1}\ln\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{f}(y_i,\boldsymbol{A}^\top\boldsymbol{x}_i)}{\widehat{f}(y_i)\widehat{f}(\boldsymbol{A}^\top\boldsymbol{x}_i)}\right\}^{\alpha-1}\right]$$
$$\approx \frac{1}{n}\sum_{i=1}^{n}\widehat{w}_{(i,\alpha)}\ln\left\{\frac{\widehat{f}(y_i,\boldsymbol{A}^\top\boldsymbol{x}_i)}{\widehat{f}(y_i)\widehat{f}(\boldsymbol{A}^\top\boldsymbol{x}_i)}\right\},$$

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

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

When $\alpha = 1$, $\widehat{w}_{(i,\alpha)} = 1$, so the maximizer of $\widehat{\mathcal{R}}_\alpha(\boldsymbol{A})$ will be essentially same as that of $\widehat{\mathcal{D}}_{KL}(\boldsymbol{A}) = \frac{1}{n}\sum_{i=1}^{n}\ln\left\{\frac{\widehat{f}(y_i,\boldsymbol{A}^\top\boldsymbol{x}_i)}{\widehat{f}(y_i)\widehat{f}(\boldsymbol{A}^\top\boldsymbol{x}_i)}\right\}$ [Yin and Cook].

Influence Function

For a fixed $\alpha \in (0,1)$, let $(Y, \boldsymbol{A}^\top \boldsymbol{X}) \sim F$. Then, our maximization problem can be considered in terms of the functional T defined as

$$T(F) = \arg\max \mathcal{R}_\alpha(\boldsymbol{A}^*) = \boldsymbol{A},$$

Let $\boldsymbol{W} = (Y, \boldsymbol{X})$ and define the contamination distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon \Delta_{\boldsymbol{w}_0}$, $0 < \varepsilon < 1$, where $\Delta_{\boldsymbol{w}_0}$ is the Dirac distribution which gives mass 1 to $\boldsymbol{w}_0 = (y_0, \boldsymbol{x}_0)$, allowing contamination of both the response and predictor vector. The influence function for T evaluated at F in the direction $\boldsymbol{w}_0$ is then defined as

$$\mathrm{IF}(T, F; \boldsymbol{w}_0) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{\partial}{\partial \epsilon} T(F_\varepsilon) \Big|_{\varepsilon=0}, \qquad (0.1)$$

and describes the effect of an infinitesimal amount of contamination at $\boldsymbol{w}_0$ on the functional T. We derive a theoretical expression for IF.

Empirical Sample Influence Function

- Let $\widehat{F} = \frac{1}{n}\sum_{j=1}^{n} \Delta_{\boldsymbol{w}_i}$ be the empirical distribution function of random sample $\{\boldsymbol{w}_i = (y_i, \boldsymbol{x}_i),\ i = 1, \cdots, n\}$ from $\boldsymbol{W} = (Y, \boldsymbol{X})$.
- Let $T(\widehat{F}) = \arg\max \widehat{\mathcal{R}}_\alpha(\boldsymbol{A}^*)$.
- The Empirical Sample Influence Function (ESIF) for T evaluated at $\widehat{F}$ in the direction of the i^{th} observation $\boldsymbol{w}_i$ is defined as

$$\text{ESIF}(T, \widehat{F}, \boldsymbol{w}_i) = \frac{\left\{T\left(\widehat{F}\right) - T\left(\widehat{F}_{(i)}\right)\right\}}{\frac{1}{(n-1)}}$$

  where $\widehat{F}_{(i)} = \{1 + (n-1)^{-1}\}\widehat{F} - (n-1)^{-1}\Delta_{\boldsymbol{w}_i}$ is the empirical distribution function with the i^{th} observation removed.
- **The ESIF quantifies the influence of each observation through the change in the estimated basis when the observation is removed.**

Sample Influence Function

Prendergast (2006) suggested Sample Influence Function (SIF) defined as

$$\mathrm{SIF}\left(\rho_{BC}, \widehat{F}, \boldsymbol{w}_i\right) = (n-1)\left\{\rho_{BC}(\widehat{\boldsymbol{A}}_{(i),k}, \widehat{\boldsymbol{A}}_k) - 1\right\},$$

where $\widehat{\boldsymbol{A}}_k = T(\widehat{F}) = [\widehat{\boldsymbol{a}}_1\ \widehat{\boldsymbol{a}}_2 \cdots \widehat{\boldsymbol{a}}_k]$ and $\widehat{\boldsymbol{A}}_{(i),k} = T(\widehat{F}_{(i)})$. The ρ_{BC} term is the Bénasséni Coefficient (BC) defined as

$$\begin{aligned}\rho_{BC}(\widehat{\boldsymbol{A}}_{(i),k}, \widehat{\boldsymbol{A}}_k) &= 1 - \frac{1}{k}\sum_{l=1}^{k} \left|\left|\widehat{\boldsymbol{a}}_l - P_{\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})}\ \widehat{\boldsymbol{a}}_l\right|\right|_2 \\ &= 1 - \frac{1}{k}\sum_{l=1}^{k} \left|\left|\left\{\mathcal{I} - P_{\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})}\right\}\ \widehat{\boldsymbol{a}}_l\right|\right|_2 \leq 1,\end{aligned}$$

where $||\cdot||_2$ is the standard matrix 2-norm, and $P_{\mathcal{S}(\widehat{\boldsymbol{A}}_{(i)})} = \widehat{\boldsymbol{A}}_{(i),k}\widehat{\boldsymbol{A}}_{(i),k}^{\top}$ is the unique orthogonal projection matrix onto $\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})$

Sample Influence Function . . .

Recall

$$\mathrm{SIF}\left(\rho_{BC}, \widehat{F}, \boldsymbol{w}_i\right) = (n-1)\left\{\rho_{BC}(\widehat{\boldsymbol{A}}_{(i),k}, \widehat{\boldsymbol{A}}_k) - 1\right\}.$$

- It is assumed that $\widehat{\boldsymbol{A}}_k$ and $\widehat{\boldsymbol{A}}_{(i),k}$ are orthonormal bases for $\mathcal{S}(\widehat{\boldsymbol{A}}_k)$ and $\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})$, respectively.
- Note that, if $\mathcal{S}(\widehat{\boldsymbol{A}}_k) = \mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})$, then $\{\mathcal{I} - P_{\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})}\}$ projects $\widehat{\boldsymbol{a}}_l$ onto $\mathcal{S}^{\perp}(\widehat{\boldsymbol{A}}_{(i),k})$.
- Consequently, $\rho_{BC}(\widehat{\boldsymbol{A}}_{(i),k}, \widehat{\boldsymbol{A}}_k) = 1$ since $\{\mathcal{I} - P_{\mathcal{S}(\widehat{\boldsymbol{A}}_{(i),k})}\}\ \widehat{\boldsymbol{a}}_l = 0$ for all $l = 1, 2, \ldots, k$.

EIF vs SIF for $y = x_1 + x_2 + x_3 + x_4 + \varepsilon$; Prendergast '06

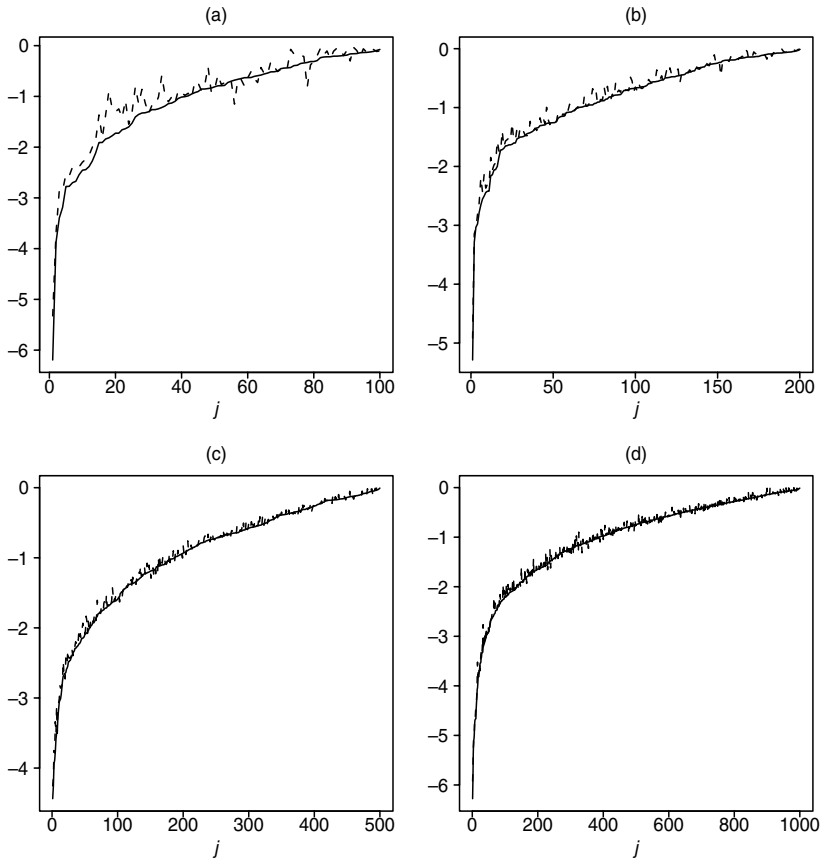

Figure: **This is for SIR method and $\varepsilon \sim N(0.1)$.** Comparison of sample influence function (SIF, solid line) and empirical influence function (EIF, dashed line) for n observations generated from the model in (6) when (a) $n = 400$, (b) $n = 600$, (c) $n = 1000$ and (d) $n = 2000$. The j-th largest $|SIF|$ is indexed by j.

EIF vs SIF for $y = \frac{x_1}{[0.5+(x_2+1.5)^2]} + 0.5\varepsilon$; Prendergast '06

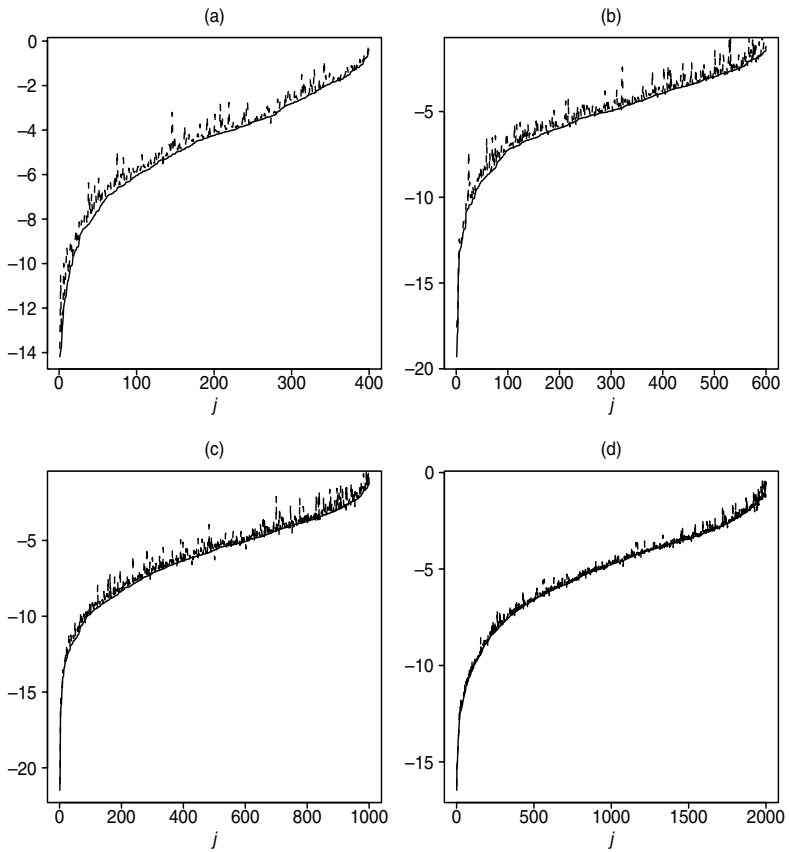

Figure: **This is for SIR method and $\varepsilon \sim N(0.1)$.** Comparison of sample influence function (SIF, solid line) and empirical influence function (EIF, dashed line) for n observations generated from the model in (5) when (a) $n = 100$, (b) $n = 200$, (c) $n = 500$ and (d) $n = 1000$. The j-th largest $|SIF|$ is indexed by j.

Accuracy Measures for Performance Assessment of $\widehat{A}$

- Two measures between the estimated and true coefficient matrices $\widehat{\mathrm{A}}$ and A are used to quantify the accuracy of the estimated basis of $\mathcal{S}_{Y|\boldsymbol{X}}$.
- The first accuracy measure is an L_2 norm defined as

$$L_{2(D)}(\widehat{\boldsymbol{A}}, \boldsymbol{A}) = \left|\left|\widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}^{\top} - \boldsymbol{A}\boldsymbol{A}^{\top}\right|\right|_2 = \left|\left|P_{\mathcal{S}(\boldsymbol{A})} - P_{\mathcal{S}(\widehat{\boldsymbol{A}})}\right|\right|_2.$$

- Note that $0 \leq L_{2(D)} \leq 2$
- The second measure of accuracy is the correlation between $\mathcal{S}(\mathrm{A})$ and $\mathcal{S}(\widehat{\mathrm{A}})$ using the square root of Hotelling's squared vector correlation coefficient,

$$\begin{aligned}\rho_{HC}(\widehat{\boldsymbol{A}}, \boldsymbol{A}) &= \sqrt{\left|(\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{T}\widehat{\boldsymbol{A}}(\widehat{\boldsymbol{A}}^{\top}\widehat{\boldsymbol{A}})^{-1}\widehat{\boldsymbol{A}}^{\top}\boldsymbol{A}\right|} \\ &= \sqrt{\left|\boldsymbol{A}^{\top}\widehat{\boldsymbol{A}}\widehat{\boldsymbol{A}}^{\top}\boldsymbol{A}\right|} = \left(\prod_{i=1}^{\mathrm{d}} \lambda_i\right)^{\frac{1}{2}}.\end{aligned}$$

Simulation Studies 1 & 2

Study 1:

- Model: $Y = \boldsymbol{a}_1^\top \boldsymbol{X}(\boldsymbol{a}_2^\top \boldsymbol{X} + 1) + \varepsilon$
- Here, $p = 10$ and $\boldsymbol{A} = [(1, 0, \ldots, 0)^\top; (0, 1, 0, \ldots, 0)^\top]$.
- $X_1 \sim t(25)$, $X_2, X_3 \sim t(5)$, $X_4, X_5 \sim N(0, 1)$, $X_6 \sim \Gamma(4, 1)$, $X_7 \sim N(0, 1)$, $X_8 \sim \chi^2_{(3)}$, $X_9 \sim \Gamma(3, 2)$, $X_{10} \sim N(0, 1)$
- $\varepsilon \sim \pi\ N(0, \sigma = .3) + (1 - \pi)\ U(0, 20)$, $\pi \in \{.95, .90\}$.

Study 2:

- Model: $Y = \frac{\boldsymbol{a}_1^\top \boldsymbol{X}}{0.5 + \left(\boldsymbol{a}_2^\top \boldsymbol{X} + 1.5\right)^2} + \varepsilon$
- Here, $p = 10$ and $\boldsymbol{A} = [(1, 0, \ldots, 0)^\top; (0, 1, 0, \ldots, 0)^\top]$.
- $X_1 \sim \Gamma(4, 3)$, $X_2 \sim t(15)$, $X_3 \sim N(0, 1)$, $X_4 \sim \chi^2_{(3)}$, $X_5 \sim t(20)$, $X_6 \sim t(25)$, $X_7 \sim N(0, 1)$, $X_8 \sim \Gamma(10, 2)$, $X_9 \sim \chi^2_{(6)}$, $X_{10} \sim N(0, 1)$
- $\varepsilon \sim \pi\ N(0, \sigma = .3) + (1 - \pi)\ U(0, 20)$, $\pi \in \{.95, .90\}$.

Plots of a simulated dataset from Studies 1 & 2

A randomly selected simulated dataset from each of Studies 1 and 2 is plotted in Figure

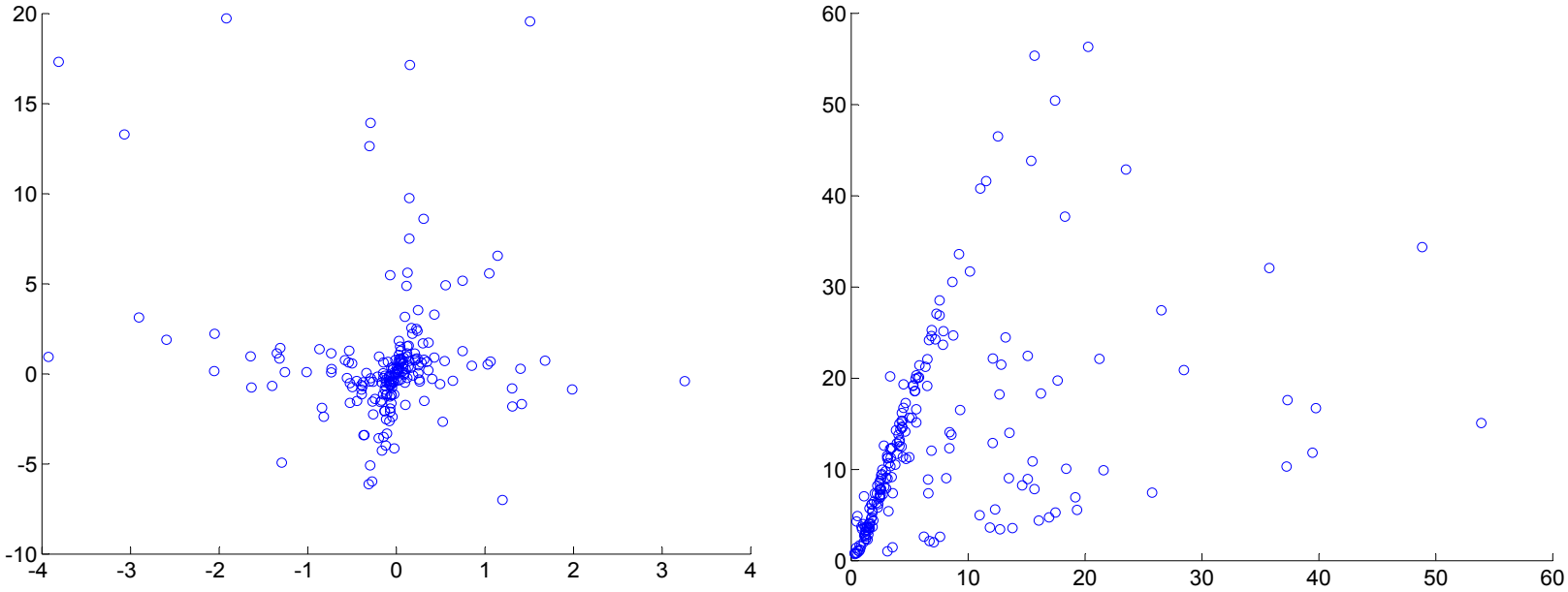

Figure: $n = 200$, $\pi = .95$. Data y versus $\boldsymbol{A}^\top \boldsymbol{x}$ plots. Left panel: Study 1, y versus $\boldsymbol{a}_1^\top \boldsymbol{x}(\boldsymbol{a}_2^\top \boldsymbol{x} + 1)$. Right panel: Study 2, y versus $\boldsymbol{a}_1^\top \boldsymbol{x}/\{0.5 + (\boldsymbol{a}_2^\top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

Results of Simulation Study 1

				Study 1					
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: Mean distance and correlations $\overline{L}_{2(D)}$ and $\overline{\rho}_{HC}$.

Determination of best α

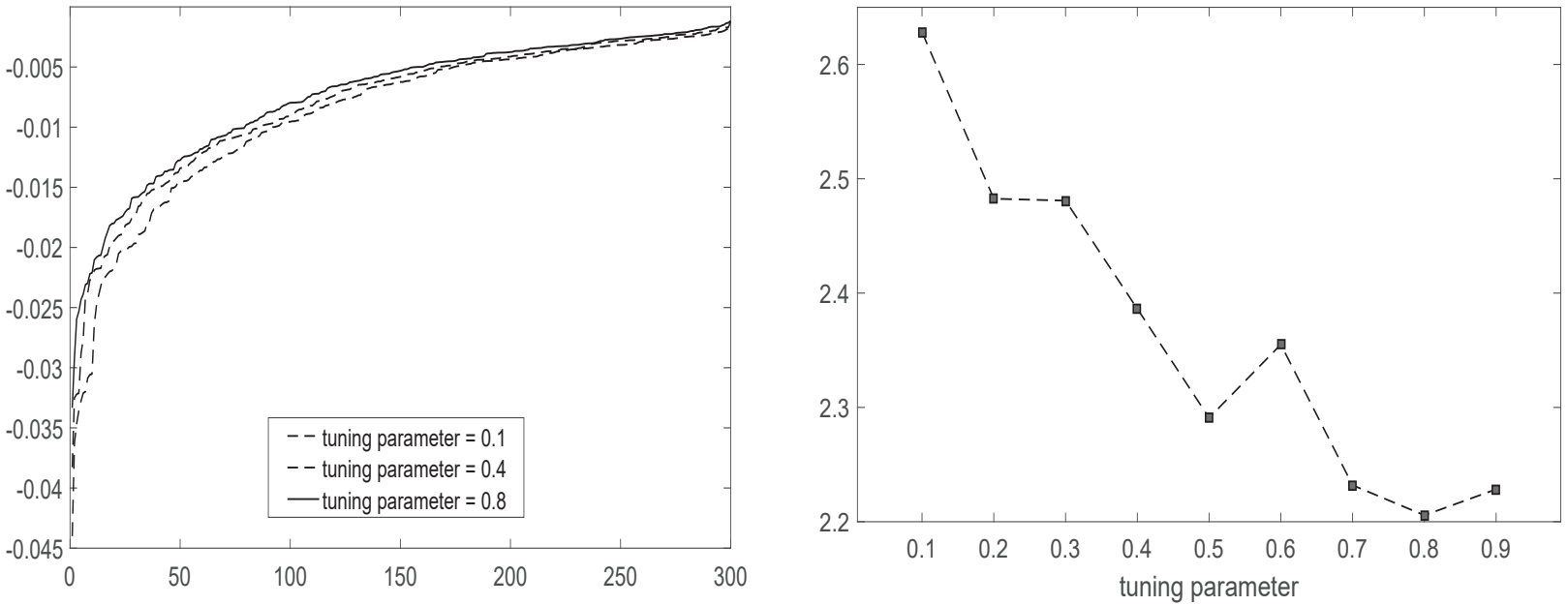

Figure: $n = 300, \pi = .90,\ \alpha = 0.8$. Left panel: smoothed SIF value plots.

Right panel: AUC_α values, $\alpha = 0.1, 0.2, \ldots, 0.9$.

Determination of Structural Dimension d

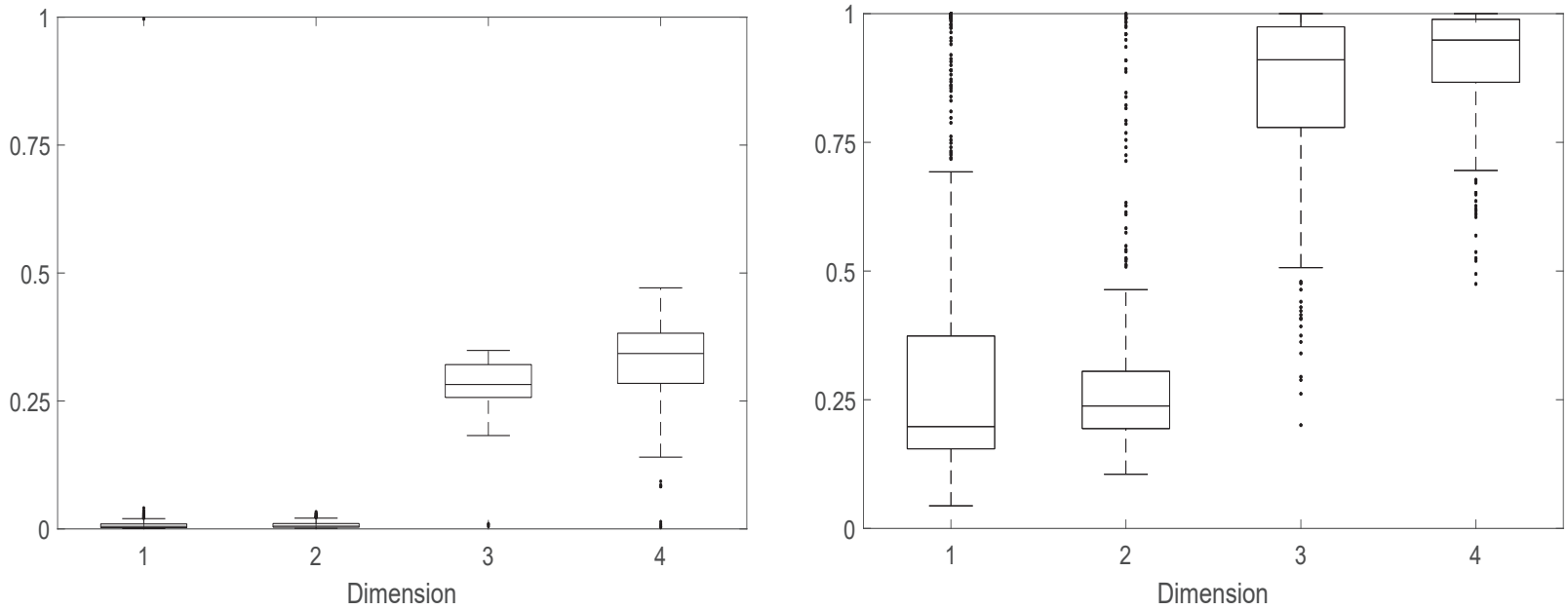

Figure: $n = 300, \pi = .90,\ \alpha = 0.8$. Left panel: boxplots of $|\mathrm{SIF}|$ values $|s_{(i,k)}|, k = 1, 2, 3, 4$. Right panel: boxplots of bootstrap $L_{2(O)}(\widehat{\boldsymbol{A}}_k^b, \widehat{\boldsymbol{A}}_k)$ values, $k = 1, 2, 3, 4$.

Results of Simulation Study 3

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

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

Determination of best α

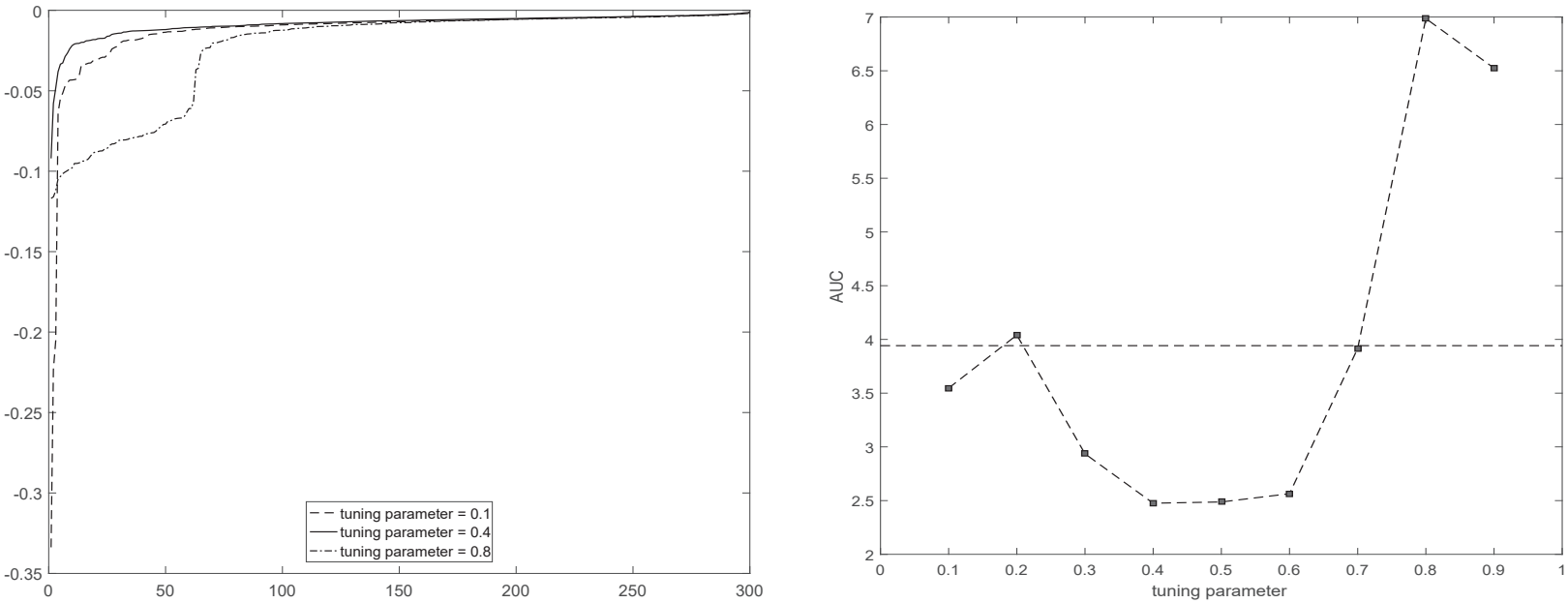

Figure: $n = 300, \pi = .90$, $\alpha = 0.4$. Left panel: smoothed SIF value plots.

Right panel: AUC_α values, $\alpha = 0.1, 0.2, \ldots, 0.9$.

Determination of Structural Dimension d

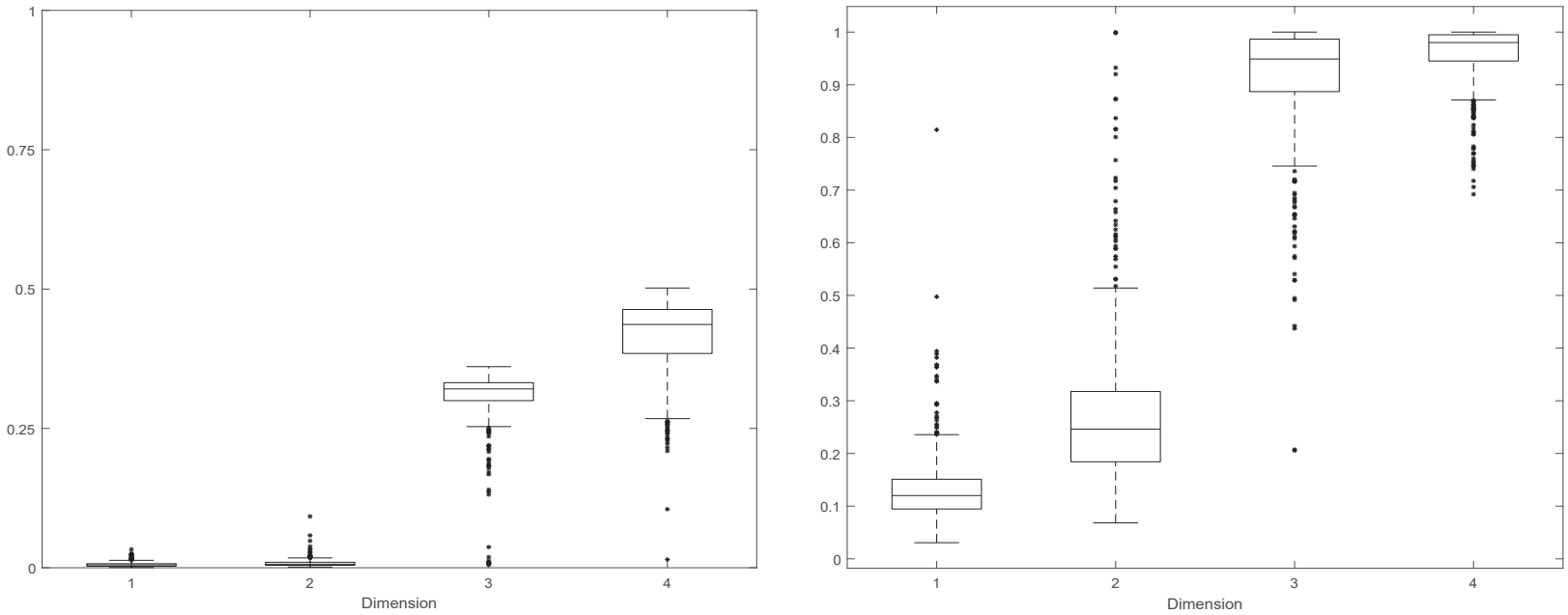

Figure: $n = 300, \pi = .90,\ \alpha = 0.4$. Left panel: boxplots of $|\mathrm{SIF}|$ values $|s_{(i,k)}|, k = 1, 2, 3, 4$. Right panel: bootstrap boxplots of $L_{2(O)}(\widehat{\mathrm{A}}_k^b, \widehat{\mathrm{A}}_k)$ values, $k = 1, 2, 3, 4$.

Hitters 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?"
- Here, the dependent variable Y is the annual salary in 1986 in natural log scale.The random vector for predicting annual salary, $\boldsymbol{X} = (X_1, X_2, \ldots, X_{16})^\top$, 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.

Hitters Salary Data Analysis

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, to improve the results they first identified outliers, and removed the observations that were deemed influential.

Hitters Salary Data Analysis

Hitter Data Analysis – $\mathcal{R}_{0.1}(\boldsymbol{A})$

	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}
$\widehat{a}_1$	.041	**.761**	.008	.101	.046	.090	-.046	.035	-.006	**.958**	.032	.064	-.019	.048	.095	.091
$\widehat{a}_2$	-.093	.125	-.010	.154	.046	-.021	-.021	.005	.001	-.013	**-.575**	-.185	**-.424**	**-.594**	.095	.208

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

Hitters Salary Data Analysis

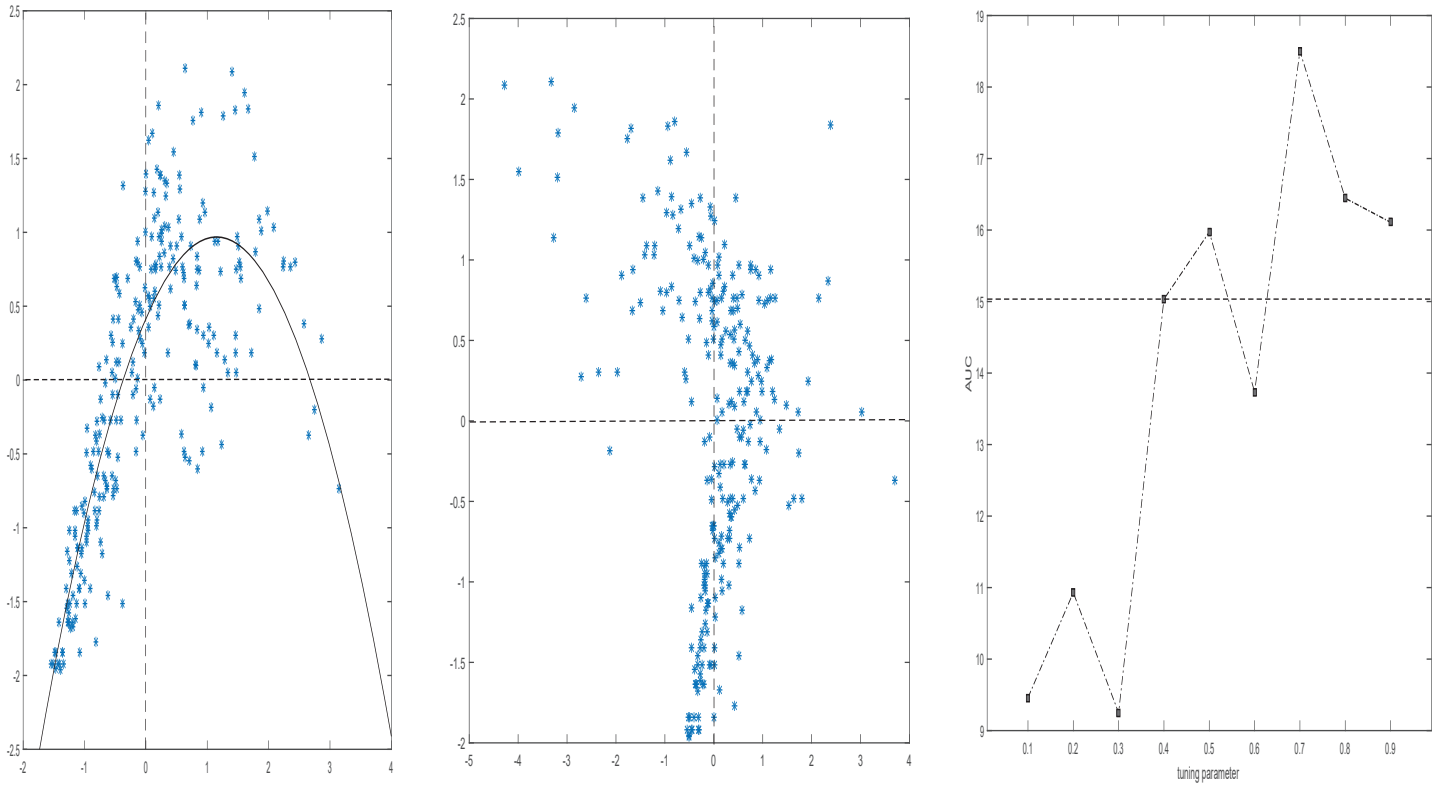

Figure: Right panel: $\widehat{\boldsymbol{a}}_1^\top \boldsymbol{x}$ vs. y, $\alpha = 0.1$. Middle panel: $\widehat{\boldsymbol{a}}_2^\top \boldsymbol{x}$ vs. y, $\alpha = 0.1$.

Right panel: AUC_α values, dimension $k = 1$, $\alpha = 0.1, 0.2, \ldots, 0.9$

Hitters Salary Data Analysis

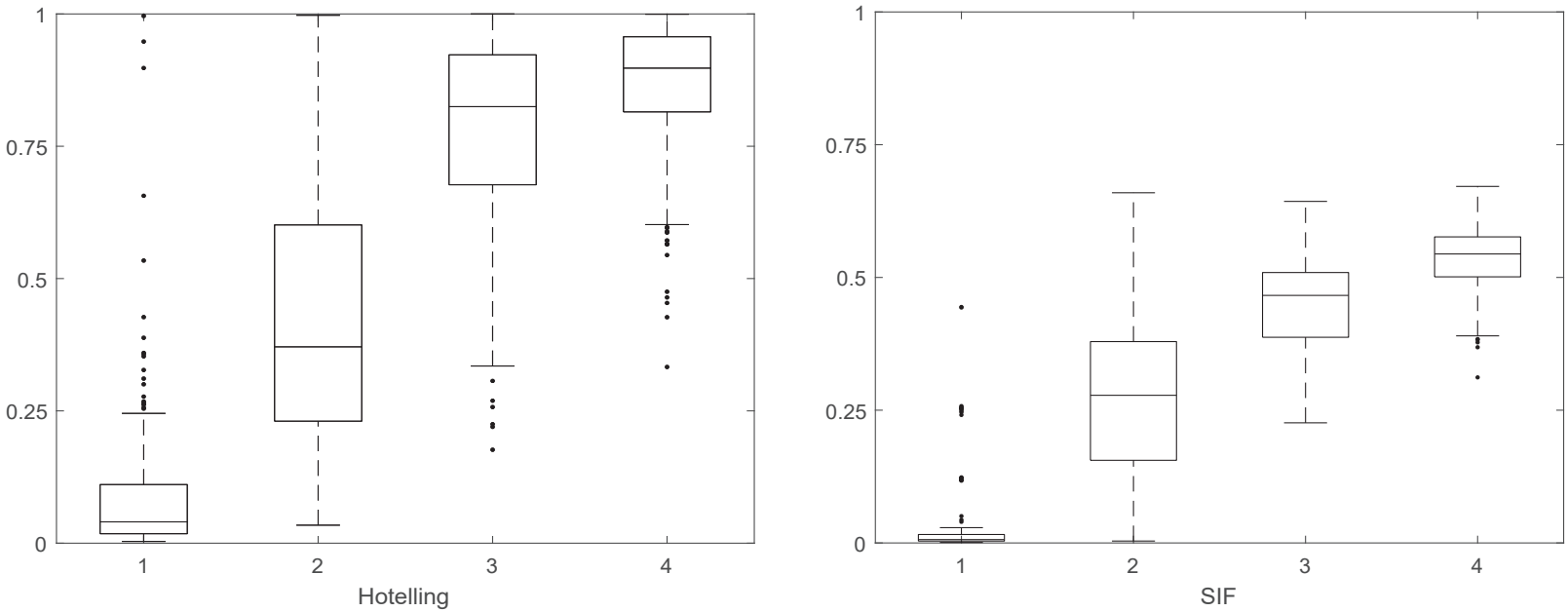

Figure: Boxplots, $\alpha = 0.1$, dimension $k = 1, 2, 3, 4$. Left Panel: Bootstrap $1 - \rho_{HC}\left(\widehat{\boldsymbol{A}}_k^b, \widehat{\boldsymbol{A}}_k\right)$ values. Right panel: $|\mathrm{SIF}|$ values $|s_{(i,k)}|$.

Hitters Salary Data Analysis

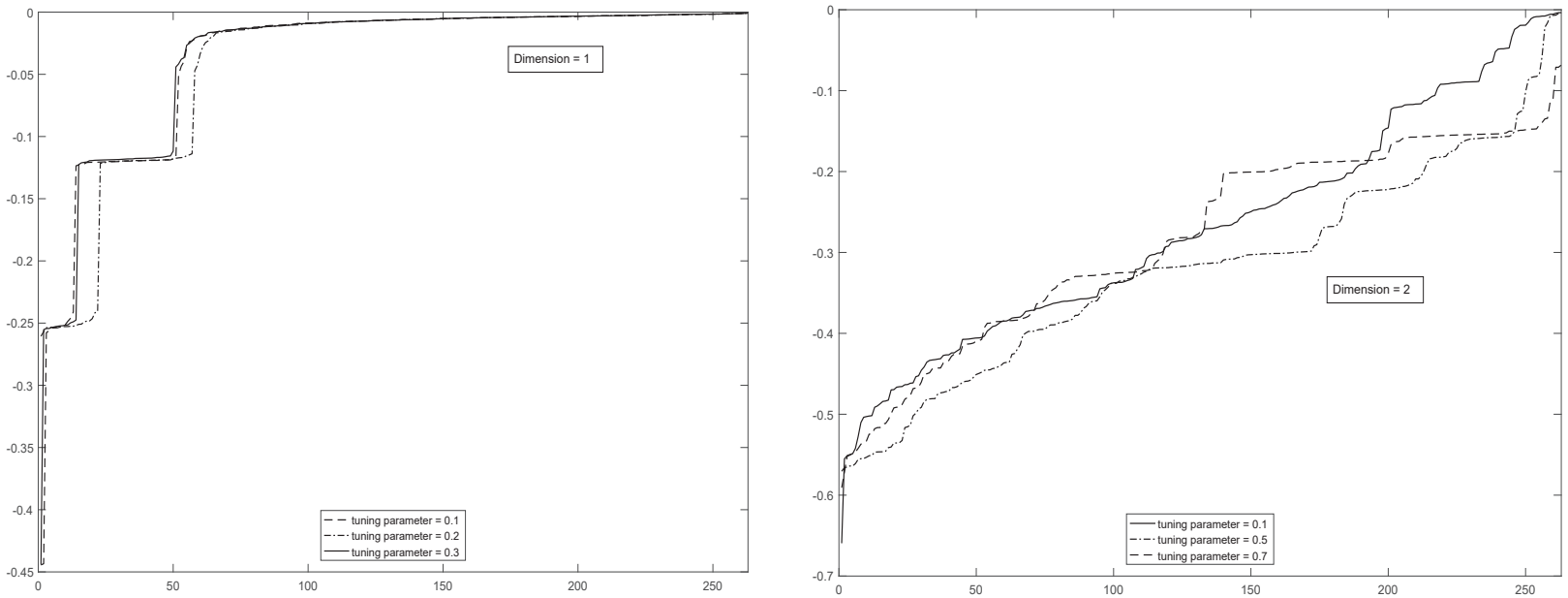

Figure: Smoothed SIF value plots. Left panel: dimension $k = 1$, $\alpha = 0.1, 0.2, 0.3$. The step function nature of the SIF plots in Figure for k $= 1$ indicates that about 50 of the n $= 263$ observations are the most influential. Right panel: dimension $k = 2$, $\alpha = 0.1, 0.5, 0.7$. (Baseball salary data analysis)

Hitters Salary Data Analysis

- As in Xia et al.(2002), after determining the two variates, $\widehat{\boldsymbol{a}}_1^\top \boldsymbol{x}$ and $\widehat{\boldsymbol{a}}_2^\top \boldsymbol{x}$, we fit a linear model using the two variates as predictors with stepwise linear regression producing the fitted model
$\widehat{y} = 0.42672 + 0.96824(\widehat{\boldsymbol{a}}_1^\top \boldsymbol{x}) - 0.228(\widehat{\boldsymbol{a}}_2^\top \boldsymbol{x}) - .42835(\widehat{\boldsymbol{a}}_1^\top \boldsymbol{x})^2$.
- 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.