Statistica Sinica Preprint No: SS-2022-0351

Title	Robust Recovery of the Central Subspace for Regression Using the Influence Function of the Rényi Divergence
Manuscript ID	SS-2022-0351
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0351
Complete List of Authors	Ross Iaci and T. N. Sriram
Corresponding Authors	Ross Iaci
E-mails	riaci@wm.edu
Notice: Accepted version subject to English editing.	

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

Ross Iaci* and T. N. Sriram

Abstract

A considerable amount of research in the literature has focused on quantifying the effect of extreme observations on classical methods for estimating the Central Subspace (CS) for regression through the study of influence functions and their sample estimates. Alternatively, a method that is inherently robust to data contamination is also important and desirable for the increased reliability in the estimation of the CS without relying on the identification and removal of influential values. To this end, we develop a new method that is innately resistant to outlying observations in recovering a dimension reduction subspace for regression based on the Rényi divergence. In addition to deriving the theoretical Influence Function (IF), the Sample Influence Function (SIF) values are directly utilized to provide new powerful and efficient methods for both estimating the dimension of the CS and selecting an optimal level of the tuning parameter to decrease the impact of extreme observations. The model-free approach is detailed theoretically, its performance investigated through simulation, and the application in practice is demonstrated through a real data analysis.

Key Words and Phrases: Bootstrap dimension estimation; Robustness; Sample Influence Function; Sufficient dimension reduction; Tuning parameter estimation.

*Ross Iaci, Department of Mathematics, The College of William and Mary, Williamsburg, VA, 23185. T.N. Sriram, Department of Statistics, University of Georgia, Athens, GA, 30602. E-mail: riaci@wm.edu, tn@stat.uga.edu.

1 Introduction

In regression analysis, the relationship between a response variable Y and a vector of explanatory variables $\mathbf{X} = (X_1, \ldots, X_p)^\top$ is the primary focus, and while the *curse of dimensionality* is a hindrance for large p, the dimension of $\mathbf{X}$ is often only artificially high since there often exists a functional relationship between Y and some lower-dimensional projection of $\mathbf{X}$. Therefore, it is natural to first identify this lower-dimensional subspace, termed a Dimension Reduction Subspace (DRS). Identification of a DRS is an important initial phase in a regression data analysis, as it not only serves as a basis, but also guides subsequent analysis when a parsimonious parametric model is not yet available.

Inspired by the pioneering methods of Sliced Inverse Regression and Sliced Average Variance Estimation (SAVE), there has been a proliferation of powerful model-free dimension reduction methods over three decades; see for example, Cook [8], Yin and Cook [38], and Iaci et al. [22], and references therein. Additionally, Sufficient Dimension Reduction (SDR) methods have been developed in Wang et al. [32] using the Hellinger integral, in Xue et al. [34] using the Hilbert-Schmidt independence criterion, and a unified approach established in Xue et al. [35] through a generalized index. Zhang et al. [41] developed a new geometric framework to reformulate the SDR problem and introduced a new concept called the Maximum Separation Subspace (MASES). They focused on the MASES under the squared Hellinger distance and developed an estimation procedure to obtain an SDR in regression and linear discriminant analysis, where the response is categorical. However, different from this article, these developed methodologies based on the Hellinger distance do not study the robustness of the respective estimation procedures under contamination. Note that, the Rényi divergence provides a general framework for robust estimation, which includes the Hellinger-Bhattacharya distance as a special case.

While useful, many Dimension Reduction (DR) methods are highly sensitive to influential observations. One way to address this issue is to study the sensitivity of the existing DR methods to extreme observations and then construct more robust versions. For example, Gather et. al [15], [14] studied the outlier sensitivity of Sliced Inverse Regression (SIR) and proposed a robust version of this method. Yohai and Noste [37] proposed another robust version of SIR by assuming that the observations in each slice have a multivariate normal distribution. More recently, Dong et al. [12] developed robust methods for SDR, termed robust inverse regression and inverse median estimation, while Zhang and Chen [40] utilized ball covariance for this objective. The Minimum Average Variance Estimation (MAVE) method of Xia et al. [33] is a popular approach for SDR, but it is also not robust against outliers in the response variable Y, as discussed in Rousseeuw and Leroy [27], Čížek [4], Critchley [11], and Čížek and Härdle [3]. To achieve robust estimation, Zhang et al. [42] developed a regularized MAVE under a nonconvex penalized regression framework, and performed a comparative study with other robust versions of MAVE.

The lack of robustness of DR methods are exacerbated for high dimensional datasets, where it is not only difficult to detect outlying and/or influential observations but often hard to resolve when they are identified. In a series of articles, Prendergast, and Prendergast and Smith (see Prendergast and Smith [25] and references therein), have derived the influence functions, and extensively studied the sample influence functions, for the effective dimension reduction directions corresponding to the SIR, principal Hessian directions (pHd) and SAV methods; Critchley [10] investigated the empirical influence functions for these methods. While these studies provide a formal way of assessing the influence of extreme observations on the estimates provided by SIR, pHd and SAVE, they do not provide a way to construct estimates that are inherently robust to data contamination.

Our goal is to provide a comprehensive methodology, based on the Rényi divergence, that recovers the intersection of all dimension reduction subspaces, termed the Central Subspace (CS), that is inherently robust to data contamination. Importantly, use of this divergence measure not only allows for the identification of both linear and nonlinear relationships between the response and a linear combination of the predictors without any model specification, but also enables us to strike a balance between efficiency and robustness against data contamination. In addition to the robust estimation of the regression dimension reduction directions, the Sample Influence Function (SIF) values can be used to determine the minimal number of significant dimensions in order to provide a sufficient dimension reduction, and to select a key index that balances efficiency and robustness, thereby providing a comprehensive approach to robust high-dimensional data analysis in regression.

The article is organized as follows. Section 1.1 introduces the Rényi divergence between two densities and gives a general motivation for its use in the robust recovery of the CS. Section 2.1 discusses the CS and its properties, including the subsequently provided minimal sufficient dimension reduction, with the developed Rényi divergence method for identifying the CS and the associated fundamental properties detailed in Section 2.2. The sample version is defined in Section 2.3, with subsequent subsections containing a heuristic argument for robustness, a consistency theorem, and two different computational algorithms. A formal assessment of the robustness is achieved through the study of the influence and estimated influence functions in Section 3. Multiple procedures for the estimation of the dimension of the CS, or the structural dimension of the regression, and the determination of the optimal level of the tuning parameter are given in Section 4. The methods developed in Sections 4.3 and 4.4 show that the calculated SIF values can be used to determine both the estimated dimension and the optimal value of the tuning parameter. The numerical studies to quantify the accuracy of the estimated central subspace are outlined in Section 5. Analysis of a baseball salary dataset, originally the focus of a sponsored section of the American Statistical Association, is revisited in Section 6 due to the known presence of outliers. The numerical studies and data analysis are carried out in detail in a Web Appendix, with an additional simulation study comparing the performance of our method to those of Zhang et al. [42].

1.1 Motivation

The Kullback Leibler (KL) divergence has been the basis of many dimension reduction methods in regression and also extended to reduce the dimensions of multiple sets of random vectors; see for example Yin and Cook [38], Iaci et al. [22], and references therein. For $\alpha > 0$, $\alpha \neq 1$, Rényi [26] defined the generalization of the KL divergence between two probability density functions $f_1(\mathbf{u})$ and $f_2(\mathbf{u})$, where $\mathbf{u}$ is a random vector. Specifically, the Rényi divergence can be defined as

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

Erven and Harremoës [13] systematically present many of the properties of (1) such as the monotonicity, continuity, and skew symmetry, as a function of α. In addition, the density divergence in (1) has the two fundamental properties that $D_\alpha\{f_1(\mathbf{U})||f_2(\mathbf{U})\} \geq 0$ for all $f_1(\mathbf{u})$ and $f_2(\mathbf{u})$, and $D_\alpha\{f_1(\mathbf{U})||f_2(\mathbf{U})\} = 0$ if and only if $f_1(\mathbf{u}) = f_2(\mathbf{u})$. There are two interesting cases for $\alpha \in (0,1)$ that provide a further motivation for the use of this divergence measure for dimension reduction. First, as a consequence of a limiting result, the Rényi divergence is bounded above by the KL divergence, that is, $\lim_{\alpha\to 1} D_\alpha\{f_1(\mathbf{U})||f_2(\mathbf{U})\} = \mathrm{E}\big[\ln\{f_1(\mathbf{U})/f_2(\mathbf{U})\}\big] = D_{KL}\{f_1(\mathbf{U})||f_2(\mathbf{U})\}$ and thus, $D_\alpha\{f_1(\mathbf{U})\ ||f_2(\mathbf{U})\} \leq D_{KL}\{f_1(\mathbf{U})|| f_2(\mathbf{U})\}$. Second, when $\alpha = 1/2$, $D_{1/2}\{f_1(\mathbf{U})||f_2(\mathbf{U})\} = -2\ \ln\big[1 - \big\{(HB)^2/2\big\}\big]$, where $HB = \Big[\int_{\mathbf{u}}\big\{\sqrt{f_1(\mathbf{u})} - \sqrt{f_2(\mathbf{u})}\big\}^2 d\mathbf{u}\Big]^{1/2}$ is the Hellinger-Bhattacharyya (HB) distance.

Another important motivation for considering this divergence is the balance between efficiency and robustness that is provided by controlling the level of the tuning parameter α; see Section 2.4. For example, Simulation Study 3 considers a complicated regression relationship between the response and predictor variables with contaminated error terms. For a simulated dataset from this study, the *standardized* SIF values for each observation are calculated using our proposed Rényi divergence based index for two levels of the tuning parameter, $\alpha = 0.4$ and 0.8; these values are plotted in the left and right panels of Figure 1, respectively. Note that, only a few observations have SIF values less than -0.05 when $\alpha = 0.4$, while a significantly larger proportion are less than -0.05 for $\alpha = 0.8$, which indicates that $\alpha = 0.4$ parameterizes the more robust index in recovering the regression DR directions.

The above discussion demonstrates that the Rényi divergence parameterized by $\alpha \in (0,1)$ can be used to provide a rich family of density divergences that include well known distances and consequently, provide a solid foundation for the development of a robust dimension reduction method for regression. Moreover, the developed methods to identify the regression DR directions utilizing this divergence produce an estimate of a basis for the CS and thus, provide a minimum sufficient dimension reduction of the predictor vector.

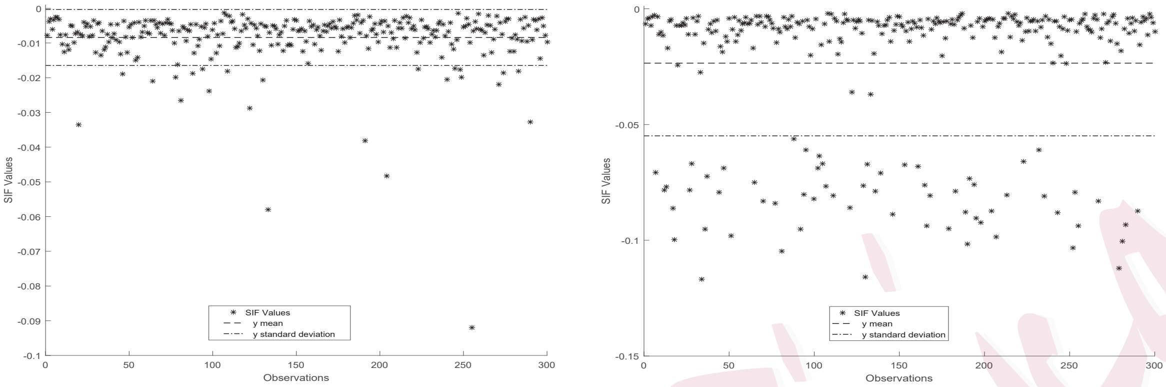

Figure 1: $n = 300, \pi = .90$: Sample influence function (SIF) values. Left panel: $\alpha = 0.4$. Right panel: $\alpha = 0.8$. (Study 3 Simulation I; successive search algorithm).

Lastly, the Rényi divergence is one of the most important and classical measures of information that is functionally connected to the λ-power divergence in Cressie and Read [9]. Both measures possess important properties that have been applied in different contexts in several disciplines. In the context of multivariate association, it is also possible, as in Iaci and Sriram [21], to use the Density Power Divergence (DPD) of Basu et al. [1]. However, the pivotal reason for using the Rényi divergence here is the ability to establish the key result in Proposition 1 part (iii) of Section 2.2 that guarantees the recovery of CS; it is unclear whether this property holds for the DPD and λ-power divergence.

2 Methodology

This section details the robust measure of association based on the Rényi density divergence that identifies both linear and nonlinear *regression* DR directions and ultimately, recovers a basis for the Central Subspace (CS).

2.1 Introduction

The goal of regression is to make an inference about the conditional distribution of a univariate response variable Y given a $p \times 1$ vector $\mathbf{X}$ of predictors. Throughout, we assume that Y and $\mathbf{X}$ are defined on a common probability space and that $(Y_i, \mathbf{X}_i)$, $i =$

$1, 2, \ldots, n$, are independent and identically distributed observations of $(Y, \mathbf{X})$ with a joint probability density function $f(y, \mathbf{x})$ and corresponding distribution function $F(y, \mathbf{x})$. The theory of sufficient dimension reduction has been developed to reduce the dimension of $\mathbf{X}$ prior to model formulation such that the full regression information between the predictors and response is preserved, while imposing the fewest probabilistic assumptions. More specifically, the goal of the dimension reduction is to identify the $p \times 1$ coefficient vectors, or directions, $\mathbf{a}_1, \ldots, \mathbf{a}_k$ such that the significant relationships between the response variable Y and the predictor vector $\mathbf{X}$ are identified through the k linear combinations $\mathbf{a}_1^\top\mathbf{X}, \ldots, \mathbf{a}_k^\top\mathbf{X}$, where $1 \leq k < p$.

To this end, let $\mathcal{S}$ denote any subspace, $\mathcal{S}(\mathbf{B})$ the k-dimensional subspace in $\mathbf{R}^p$ spanned by the columns the matrix $\mathbf{B}$, and $P_{\mathcal{S}(\mathbf{B})}$ the projection onto $\mathcal{S}(\mathbf{B})$ with respect to the usual inner product. The subspace $\mathcal{S}(\mathbf{B})$ is a Dimension Reduction Subspace (DRS) for the regression of Y on $\mathbf{X}$ if Y is conditionally independent of $\mathbf{X}$ given the projection of $\mathbf{X}$ onto $\mathcal{S}(\mathbf{B})$, denoted $Y \perp\!\!\!\perp \mathbf{X}|P_{\mathcal{S}(\mathbf{B})}\mathbf{X}$. That is, $\mathcal{S}(\mathbf{B})$ is a DRS if $f(y, \mathbf{x}|\ P_{\mathcal{S}(\mathbf{B})}\mathbf{x}) = f(y|\ P_{\mathcal{S}(\mathbf{B})}\mathbf{x})f(\mathbf{x}|\ P_{\mathcal{S}(\mathbf{B})}\mathbf{x})$, or equivalently $f(y|\ \mathbf{x}) = f(y|\ P_{\mathcal{S}(\mathbf{B})}\mathbf{x})$, for all $(y, \mathbf{x}) \in \mathbf{R} \times \mathbf{R}^p$. Importantly, the conditional independence holds if $\mathbf{B}$ is replaced with any matrix $\mathbf{B}^*$ such that $\mathcal{S}(\mathbf{B}^*) = \mathcal{S}(\mathbf{B})$, which means that any basis of a DRS is also a DRS. Also, when $\mathcal{S}(\mathbf{B})$ is a DRS, the transformation $\mathbf{B}^\top\mathbf{X}$ provides a *sufficient* dimension reduction.

Next, let $\mathcal{S}_{Y|\mathbf{X}}$ denote the intersection of all DRSs, which is a DRS under mild conditions and termed the Central Subspace (CS); see Cook [8]. The true dimension of the CS, $d = dim(\mathcal{S}_{Y|\mathbf{X}})$, commonly termed the structural dimension for the regression of Y on $\mathbf{X}$, is typically far less than the dimension p. Suppose that $\mathbf{A}$ is a $d < p$ dimensional basis for $\mathcal{S}_{Y|\mathbf{X}}$, then the conditional distributions of $Y|\mathbf{A}^\top\mathbf{X}$ and $Y|\mathbf{X}$ are the same and consequently, $\mathbf{A}^\top\mathbf{X}$ and $\mathbf{X}$ contain the same information for the regression. That is, given the *minimum* sufficient dimension reduction $\mathbf{A}^\top\mathbf{X}$, the remaining feature of $\mathbf{X}$ can be discarded without sacrificing predictive power. Throughout, we assume that $\mathcal{S}_{Y|\mathbf{X}}$ exists with structural dimension d and focus on the robust estimation of a basis $\mathbf{A}$ for $\mathcal{S}_{Y|\mathbf{X}}$.

2.2 Robust identification of the CS

Consider the $p \times 1$ random vector $\mathbf{X}$, response variable Y, and $p \times k$ matrix $\mathbf{A}$ with $k \leq p$. Next, let $f(Y, \mathbf{A}^\top\mathbf{X})$, $f(\mathbf{A}^\top\mathbf{X})$ and $f(Y)$ denote the joint and marginal densities of Y and the linear transformation $\mathbf{A}^\top\mathbf{X}$. To recover $\mathcal{S}_{Y|\mathbf{X}}$ and provide a robust method for dimension reduction in regression, for each $\alpha \in (0,1)$ consider a new Rényi divergence-based index, denoted $\mathcal{R}_\alpha(\mathbf{A})$, defined as

$$\begin{aligned}\mathcal{R}_\alpha(\mathbf{A}) = D_\alpha\{f(Y{,}\mathbf{A}^\top\mathbf{X})||f(Y)f(\mathbf{A}^\top\mathbf{X})\} &= \frac{1}{\alpha-1}\ln\left[\mathrm{E}\left\{\frac{f(Y, \mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha-1}\right] \\ &= \frac{1}{\alpha-1}\ln\left[\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left\{\frac{f(y, \mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\right\}^{\alpha-1} f(y, \mathbf{A}^\top\mathbf{x})\, d(\mathbf{A}^\top\mathbf{x})\, dy\right]. \quad (2)\end{aligned}$$

Letting $\mathbf{U} = (Y, \mathbf{A}^\top\mathbf{X})$ in (1), then $D_\alpha\{f_1(Y, \mathbf{A}^\top\mathbf{X})||f_2(Y, \mathbf{A}^\top\mathbf{X})\} \geq 0$ and $D_\alpha\{f_1(Y, \mathbf{A}^\top\mathbf{X})||\, f_2(Y, \mathbf{A}^\top\mathbf{X})\} = 0$ if and only if $f_1(y, \mathbf{A}^\top\mathbf{x}) = f_2(y, \mathbf{A}^\top\mathbf{x})$ by definition. Next, defining $f_1(y, \mathbf{A}^\top\mathbf{x}) = f(y, \mathbf{A}^\top\mathbf{x})$ and $f_2(y, \mathbf{A}^\top\mathbf{x}) = f(y)f(\mathbf{A}^\top\mathbf{x})$, which are both probability density functions, then these results hold for (2) and consequently, $\mathcal{R}_\alpha(\mathbf{A})$ is bounded below by zero with equality if and only if $Y \perp\!\!\!\perp \mathbf{A}^\top\mathbf{X}$. Also, by Proposition 1 part (iii), $\mathcal{R}_\alpha(\mathbf{A}) \leq \mathcal{R}_\alpha(\mathbf{I})$ and consequently, if $Y \perp\!\!\!\perp \mathbf{X}$ then $\mathcal{R}_\alpha(\mathbf{I}) = \mathcal{R}_\alpha(\mathbf{A}) = 0$.

As noted in Section 1.1, the limit of (2) as $\alpha \to 1$ is

$$\begin{aligned}\lim_{\alpha\to 1}\mathcal{R}_\alpha(\mathbf{A}) = D_{KL}(\mathbf{A}) &= D_{KL}\{f(Y, \mathbf{A}^\top\mathbf{X})||f(Y)f(\mathbf{A}^\top\mathbf{X})\} \\ &= \mathrm{E}\{\ln\left[f(Y, \mathbf{A}^\top\mathbf{X})/\{f(Y)f(\mathbf{A}^\top\mathbf{X})\}\right]\},\end{aligned}$$

which is the Kullback-Leibler (KL) divergence considered in the Expected Log-likelihood (EL) methods of Yin and Cook [38]. Also, for a fixed $\mathbf{A}$, due to the monotonicity property of the Rényi divergence, $\mathcal{R}_{\alpha_1}(\mathbf{A}) \leq \mathcal{R}_{\alpha_2}(\mathbf{A}) \leq D_{KL}(\mathbf{A})$ for $0 < \alpha_1 < \alpha_2 < 1$; details of these two properties are discussed in Appendix A.5 for completeness. When $\alpha = 1/2$, $\mathcal{R}_{1/2}(\mathbf{A})$ is equivalent to the Hellinger - Bhattacharyya (HB) distance and therefore, $\{\mathcal{R}_\alpha(\mathbf{A}); \alpha \in (0,1)\}$ provides a continuous, and non-decreasing, family of divergences that includes log-likelihood association, inverse regression and the KL divergence; equivalence to the HB distance is shown in Appendix A.6. The fundamental properties for recovering a basis for

$\mathcal{S}_{Y|\mathbf{X}}$ using (2) are stated in the following proposition.

Proposition 1: Let $\mathbf{A}$ and $\mathbf{A}_1$ 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 $\mathbf{I}$, then the following hold:

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

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

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

The proof of Proposition 1 is given in Appendix A.1.

Part (i) implies that searches made successively through increasing dimensional subspaces will ultimately yield a basis for $\mathcal{S}_{Y|\mathbf{X}}$ when part (iii) is satisfied. Part (ii) implies that matrices that span the same subspace have the same measured dependence and therefore, only a basis for the subspace is needed.

Important to the goals of this paper, part (iii) first establishes the bound $\mathcal{R}_\alpha(\mathbf{A}) \leq \mathcal{R}_\alpha(\mathbf{I})$, which indicates that the most dependence between $\mathbf{X}$ and Y in k dimensions can be recovered by maximizing $\mathcal{R}_\alpha(\mathbf{A})$ with respect to $\mathbf{A}$. Next, if $\mathcal{R}_\alpha(\mathbf{A}) = \mathcal{R}_\alpha(\mathbf{I})$ for a fixed α, then $\mathbf{A}$ with full rank provides a basis for a k-dimensional DRS in $\mathbf{R}^p$ and accordingly, $\mathbf{A}^\top\mathbf{X}$ is a sufficient dimension reduction for the regression of Y on $\mathbf{X}$. This also implies that when the column dimension of $\mathbf{A}$ is equal to the structural dimension of the regression, $k = d$, and the equality holds, then $\mathbf{A}$ is a basis for $\mathcal{S}_{Y|\mathbf{X}}$ and respectively, $\mathbf{A}^\top\mathbf{X}$ provides a minimum sufficient dimension reduction. That is, for d known and α fixed, a basis $\mathbf{A}_{p\times d}$ for $\mathcal{S}_{Y|\mathbf{X}}$ can be recovered as

$$\mathbf{A} = \arg\max \mathcal{R}_\alpha(\mathbf{A}^*) \quad \text{subject to the constraint} \quad \mathbf{A}^\top \Sigma_{\mathbf{X}} \mathbf{A} = \mathbf{I}, \tag{3}$$

where $\Sigma_{\mathbf{X}}$ is the covariance matrix of the explanatory vector $\mathbf{X}$. This constraint ensures that $\mathbf{A}$ is full rank and that each of the transformations $\mathbf{a}_i^\top\mathbf{X}$, termed variates, have unit variance, $\text{var}(\mathbf{a}_i^\top\mathbf{X}) = 1$ for all $i = 1, 2, \ldots, d$, and are uncorrelated with all other variates, $\text{cov}(\mathbf{a}_i^\top\mathbf{X}, \mathbf{a}_j^\top\mathbf{X}) = 0$ when $i \neq j$; these are equivalent to the well known principle component analysis constraints.

Note that, the EL divergence index of Yin and Cook [38] also satisfies Proposition 1, but that using $\mathcal{R}_\alpha(\mathbf{A})$ has a distinct advantage over their procedure in that the tuning parameter α can be exploited to naturally down-weigh outliers and thereby, yield robust estimates of a basis for $\mathcal{S}_{Y|\mathbf{X}}$; see Sections 2.4 and 3 for more details.

The next proposition shows that $\mathcal{R}_{\alpha}(\mathbf{A})$ is invariant under a full rank linear transformation of the explanatory vector and a scalar multiple of the response. For notational convenience, let $\mathcal{R}_{(U_1,\mathbf{U}_2),\alpha}(\mathbf{A})$ represent the divergence measured in the $(U_1, \mathbf{U}_2)$ scale, where U_1 is any random variable and $\mathbf{U}_2$ any $p \times 1$ random vector.

Proposition 2: Consider the arbitrary scalars C_1 and a, and any nonsingular $p \times p$ matrix $\mathbf{C}_2$ and $p \times 1$ vector $\mathbf{b}$. Then, for the transformations $W_1 = C_1^{-1}Y + a$ and $\mathbf{W}_2 = \mathbf{C}_2^{-1}\mathbf{X} + \mathbf{b}$ the following holds:

$$\mathcal{R}_{(Y,\mathbf{X}),\alpha}(\mathbf{A}) = \mathcal{R}_{(W_1,\mathbf{W}_2),\alpha}(\mathbf{C}_2^{\top}\mathbf{A}), \textit{ which implies } \mathcal{R}_{(Y,\mathbf{X}),\alpha}(\mathbf{I}) = \mathcal{R}_{(W_1,\mathbf{W}_2),\alpha}(\mathbf{C}_2^{\top}\mathbf{I}).$$

The proof of Proposition 2 is given in Appendix A.2.

Proposition 2 states that the index in (2) is invariant under linear transformations, establishing that is these types of transformations of the response and predictor vector do not affect the associations that exist between $\mathbf{X}$ and Y. Accordingly, if $\mathbf{A}_{\mathbf{w}_2}$ is a coefficient matrix in the transformed scale, then $\mathbf{A}_{\mathbf{w}_2}^{\top}\mathbf{W}_2$ is recovered in the original scale as $\mathbf{A}_{\mathbf{w}_2}^{\top}\{\mathbf{C}_2^{-1}\mathbf{X} + \mathbf{b}\} = (\mathbf{C}_2^{-\top}\mathbf{A}_{\mathbf{w}_2})^{\top}\mathbf{X} + \mathbf{A}_{\mathbf{w}_2}^{\top}\mathbf{b}$, which implies that the coefficient matrix in the original scale is $\mathbf{C}_2^{-\top}\mathbf{A}_{\mathbf{w}_2}$.

Of computational importance, discussed in more detail in Section 2.6, for the transformations $Z_Y = \sigma^{-1/2}\{Y - \mathrm{E}(Y)\}$ and $\mathbf{Z}_{\mathbf{X}} = \Sigma^{-1/2}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}$, the constraint in (3) is reduced to $\mathbf{A}^{\top}\mathbf{A} = \mathbf{I}$; note that the coefficient matrix in the original scale is $\Sigma_{\mathbf{X}}^{1/2}\mathbf{A}$. Also, for $k < p$, $\mathbf{A}$ is a semi-orthogonal matrix and therefore, an orthonormal basis for $\mathcal{S}(\mathbf{A})$. For convenience, a matrix of any dimension $k \leq p$ with orthonormal columns is termed *orthonormal*.

2.3 Sample estimation

Consider a random sample $\{(y_i, \mathbf{x}_i); i = 1, 2, \ldots, n\}$ from $(Y, \mathbf{X})$, and assume that the structural dimension d of the regression is known. Since no distributional assumptions are made, the densities in (2) are unknown and therefore, for any $p \times k$ matrix $\mathbf{A}$ and $\alpha \in (0, 1)$, we define the following sample estimate

$$\widehat{\mathcal{R}}_\alpha(\mathbf{A}) = \frac{1}{\alpha-1}\ \ln\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\widehat{f}(y_i,\mathbf{A}^\top\mathbf{x}_i)}{\widehat{f}(y_i)\widehat{f}(\mathbf{A}^\top\mathbf{x}_i)}\right\}^{\alpha-1}\right], \tag{4}$$

where $\widehat{f}(y_i)$, $\widehat{f}(\mathbf{A}^\top\mathbf{x}_i)$ and $\widehat{f}(y_i,\mathbf{A}^\top\mathbf{x}_i)$ are kernel density estimates of $f(y_i)$, $f(\mathbf{A}^\top\mathbf{x}_i)$ and $f(y_i,\mathbf{A}^\top\mathbf{x}_i)$, respectively. Specifically, to estimate $f(y_i,\mathbf{A}^\top\mathbf{x}_i)$ for a specific coefficient matrix $\mathbf{A} = [\mathbf{a}_1\ \mathbf{a}_2\cdots\mathbf{a}_k]$, we use the Gaussian product kernel density estimate

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

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 $\{\mathbf{a}_l^\top\mathbf{x}_i, i = 1,\ldots,n\}$, respectively. The above formula is modified to provide an estimate of the marginal density $f(\mathbf{A}^\top\mathbf{x}_i)$ as $\widehat{f}\big(\mathbf{A}^\top\mathbf{x}_i\big) = \frac{1}{nh_1h_2\cdots h_k}\sum_{j=1}^{n}\big(\prod_{l=1}^{k}K\big[\{\mathbf{a}_l^\top(\mathbf{x}_j-\mathbf{x}_i)\}/h_l\big]\big)$; similarly, the density estimate of $f(y_i)$ is $\widehat{f}\big(\mathbf{y}_i\big) = \frac{1}{nh^*}\sum_{j=1}^{n}\big(K\big[\{(\mathbf{y}_j-\mathbf{y}_i)\}/h^*\big]\big)$. Note that, due to the limiting result of (2) as $\alpha \to 1$, the sample estimate in (4) can defined as $\widehat{\mathcal{D}}_{KL}(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^{n}\ln\Big[\widehat{f}(y_i,\mathbf{A}^\top\mathbf{x}_i)/\{\widehat{f}(y_i)\widehat{f}(\mathbf{A}^\top\mathbf{x}_i)\}\Big]$ at $\alpha = 1$, which is the sample version of the KL based method of Yin and Cook [38].

The suggested use of Gaussian product kernels in Scott [29] and Silverman [31] were shown to work well in the simulation studies testing the performance of the KL based methods of association in Yin and Sriram [39], and Iaci et al. [22]; the bandwidth selection was also supported by the results. Successful implementations of this pairing are also referenced in Iaci et. al [22]. Additionally, Iaci et al. [20] compared the Gaussian and Epanechnikov kernels at different bandwidths in the estimation of their L_2 based measure of association and noted a slight improvement in performance using the above kernel bandwidth combination. Noting that the index in (2) provides a smooth bridge between the KL and an L_2 distance, for $0 < \alpha < 1$, further motivated our choice. The performance in the simulation studies in Section 5 support our kernel and bandwidth selection and importantly, the properties of the method hold for any kernel of bounded variation; see Theorem 1 of Section 2.5.

With the structural dimension d known, then based on the discussion of Proposition 1 in Section 2.2, an estimate of a basis for $\mathcal{S}_{Y|\mathbf{X}}$ can be recovered by maximizing the sample version of (2) with respect to the $p \times d$ matrix $\mathbf{A}$. To this end, for $\alpha \in (0,1)$, our Rényi divergence based estimator of $\mathbf{A}$ is defined as

$$\widehat{\mathbf{A}} = \operatorname{argmax}\, \widehat{\mathcal{R}}_\alpha(\mathbf{A}^*) \quad \text{subject to the constraint} \quad \widehat{\mathbf{A}}^\top \widehat{\Sigma}_{\mathbf{x}} \widehat{\mathbf{A}} = \mathbf{I}, \tag{5}$$

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

2.4 Heuristic argument for robustness

The inherent robustness of (4) and the role of the tuning parameter in balancing efficiency and robustness, is motivated through a heuristic argument in Web Appendix B since a more formal assessment is provided in Section 3.

2.5 Consistency

In this section a consistency result is stated for the estimated coefficient matrix $\widehat{\mathbf{A}}$ defined in Section 2.3 under the constraint $\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} = \mathbf{I}$, with the proof given in Appendix A.3.

First, define the set

$$\chi_b = \left\{ i : f(y_i) > b, f(\mathbf{A}^\top\mathbf{x}_i) > b,\ \text{ and } f(y_i, \mathbf{A}^\top\mathbf{x}_i) > b,\ \text{ for any } \mathbf{A} \text{ such that } \mathbf{A}^\top\mathbf{A} = \mathbf{I} \right\},$$

for some $b > 0$ given in the proof, and let n_{b^c} denote the number of observations whose indices are not in χ_b. We then have the following result.

Theorem 1 (Consistency) *Assume the conditions of Lemma* 1 *in Appendix A.3, and that $n_{b^c}/n \to 0$ as $n \to \infty$. Let $\widehat{\mathbf{A}}^b = \arg\max \widehat{\mathcal{R}}^b_\alpha(\mathbf{A}^*)$ and $\mathbf{A} = \arg\max \mathcal{R}_\alpha(\mathbf{A}^*)$, for each $\alpha \in (0,1)$, where*

$$\widehat{\mathcal{R}}^b_\alpha(\mathbf{A}^*) = \frac{1}{\alpha-1} \ln\left[\frac{1}{n}\sum_{i=1}^{n} J(i \in \chi_b) \left\{ \frac{f_n(y_i, \mathbf{A}^{*\top}\mathbf{x}_i)}{f_n(y_i) f_n(\mathbf{A}^{*\top}\mathbf{x}_i)} \right\}^{\alpha-1}\right],$$

with f_n defined in Appendix A.3 and $J\left(i \in \chi_b\right)$ is the indicator function for χ_b. Then,

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

2.6 Computational algorithms

Due to the invariance of $\mathcal{R}_\alpha(\mathbf{A})$ under nonsingular matrix transformations, the response and explanatory vector can be mean centered and transformed as $Z_Y = \sigma_Y^{-1/2}\{Y - \mathrm{E}(Y)\}$ and $\mathbf{Z}_{\mathbf{X}} = \Sigma_{\mathbf{X}}^{-1/2}\{\mathbf{X} - \mathrm{E}(\mathbf{X})\}$. These transformations change the scale, but not the relationships between the response and predictors, and removes any multicollinearity between the explanatory variables. The transformation of $\mathbf{X}$ also lessens the effect of variables of differing magnitude and importantly, as addressed following Proposition 2 in Section 2.2, the coefficient matrix in the original scale can easily be recovered as $\Sigma_{\mathbf{X}}^{1/2}\mathbf{A}$. Note that, under these transformations the sample constraint in (5) is simplified to $\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} = \mathbf{I}$ and therefore, $\widehat{\mathbf{A}}$ is an orthonormal matrix, semi-orthogonal for $k < p$ and orthogonal for $k = p$, with the columns providing an orthonormal basis for a k-dimensional subspace in $\mathbf{R}^p$. Since the estimated coefficient vectors $\widehat{\mathbf{a}}_l, l = 1, \ldots, k,\ k \leq p$, that identify the relationships are of interest, and not the scale of the regression DR directions, $\widehat{\mathbf{A}}^\top\widehat{\mathbf{A}} = \mathbf{I}$ is referred to as the orthonormal constraint. For simplicity, the algorithms are put forward assuming that the response and predictor vector are in this *whitened* scale, but maintain the notation Y and $\mathbf{X}$ for continuity.

Two methods for estimating $\widehat{\mathbf{A}}$ are considered. The first is a direct matrix maximization approach, and the second a successive search for each of the estimated coefficient vectors $\widehat{\mathbf{a}}_l, l = 1, 2, \ldots, k$. The maximization of the sample index in (4) under the orthonormal constraint is achieved using the nonlinear constrained minimizer *fmincon* in MATLAB, which uses a Sequential Quadratic Programming procedure that simultaneously incorporates the constraints.

Direct search algorithm:

This matrix maximization algorithm is a modification of the one used in Iaci et al. [22] that employed an alternating search procedure to provide a sufficient dimension reduction of two random vectors in a multivariate association setting.

Step 0: Set $l = 0$ and generate a $p \times k$ initial guess matrix $\widehat{\mathbf{A}}_0$ to be supplied to the *fmincon* function.

> – Orthonormal matrices of dimension $p \times k$ are generated at random, termed type 1 initial guesses. Next, type 2 orthogonal matrices are generated at random with columns consisting only of zeros and ones. Let $I_1 = \{\mathbf{B}_{1,i}; i = 1, 2, \ldots N_1\}$ and $I_2 = \{\mathbf{B}_{2,j}; j = 1, 2 \ldots, N_2\}$ denote the sets of N_1 type 1 and N_2 type 2 matrices, respectively. The best initial guess $\widehat{\mathbf{A}}_0$ is taken to be the matrix that generates the largest sample index value in (4) among all randomly generated matrices in $I_1 \cup I_2$.

Step 1: Find $\widehat{\mathbf{A}}_{l+1} = \arg\max \widehat{\mathcal{R}}_\alpha(\mathbf{A}^*)$, subject to the constraint $\widehat{\mathbf{A}}_{l+1}^\top \widehat{\mathbf{A}}_{l+1} = \mathbf{I}$. That is, determine $\widehat{\mathbf{A}}_{l+1}$ such that the sample index in (4) is maximum and $\widehat{\mathbf{A}}_{l+1}^T \widehat{\mathbf{A}}_{l+1} = \mathbf{I}$.

Step 2: Let $\widehat{\mathbf{A}}_{l+1}$ be the new initial guess. Increment l by 1. If $l = 1$, repeat step 1 so that there are at least two iterations for the comparison in step 3.

Step 3: Repeat steps 1 and 2 until the difference $\big[\widehat{\mathcal{R}}_\alpha(\widehat{\mathbf{A}}_l) - \widehat{\mathcal{R}}_\alpha(\widehat{\mathbf{A}}_{l-1})\big]$ is less than a user defined tolerance, say 10^{-6}, or the user defined maximum number of iterations has been reached.

The minimum number of iterations taken was four and found to be more than necessary for the difference in successive estimated index values to be within 10^{-6}. In practice, the convergence settings in the *fmincon* function can be relied on in performing steps 0 and 1 only.

Successive search algorithm:

The algorithm for estimating $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \cdots \widehat{\mathbf{a}}_k]$ by searching for the estimated coefficient vectors $\widehat{\mathbf{a}}_l, l = 1, 2, \ldots, k$, in succession is detailed here, and is a modification of the method presented in Iaci et. al [20] to estimate the coefficient vectors that recover the relationships between m-sets of random vectors.

Step 0: Set $l = 1$, generate an initial guess $\mathbf{b}$ to be supplied to the *fmincon* function and determine the coefficient vector $\widehat{\mathbf{a}}_1 = \arg\max \widehat{\mathcal{R}}_\alpha(\mathbf{a})$, subject to the constraint $\widehat{\mathbf{a}}_1^\top \widehat{\mathbf{a}}_1 = 1$. That is, find $\widehat{\mathbf{a}}_1$ such that the sample index in (4) is maximized and $\widehat{\mathbf{a}}_1^\top \widehat{\mathbf{a}}_1 = 1$.

> – An initial guess is produced by first generating a p-dimensional random vector consisting of zeros and ones, say $\mathbf{b}_j^*$, and then normalizing it to have unit length, $\mathbf{b}_j = \mathbf{b}_j^*/||\mathbf{b}_j^*||$. Let $I = \{\mathbf{b}_j; j = 1, 2, \ldots, N\}$ denote the set of N initial guesses. The best initial guess is the vector $\mathbf{b}$ that produces the largest sample index value of (4) among all randomly generated vectors in I.

Step 1: If $l = k$ then stop, else increment l by 1. Use the singular value decomposition to determine the left singular vectors of the matrix $\mathbf{A}^* = [\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \cdots \widehat{\mathbf{a}}_{l-1}][\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \cdots \widehat{\mathbf{a}}_{l-1}]^\top$. Define the matrix $\mathbf{U} = [\mathbf{u}_1\ \mathbf{u}_2 \cdots \mathbf{u}_p]$, where $\mathbf{u}_1, \ldots, \mathbf{u}_p$ are the left singular vectors of $\mathbf{A}^*$.

Step 2: Let $\mathbf{U}^* = [\mathbf{u}_k\ \mathbf{u}_{k+1} \cdots \mathbf{u}_p]$ and project the data matrix, denoted $\mathbf{D}_{n\times p}$, onto the subspace spanned by the columns of $\mathbf{U}^*$. That is, create the new data matrix $\mathbf{D}^*_{n\times\{p-(l-1)\}} = \mathbf{D}\mathbf{U}^*$.

Step 3: Based on the data matrix $\mathbf{D}^*$, generate an initial guess as detailed in *Step* 0 and determine the $\{p-(l-1)\}\times 1$ coefficient vector $\widehat{\mathbf{a}}_l^*$ that maximizes the index in (4) subject to the constraint $\widehat{\mathbf{a}}_l^{*\top}\widehat{\mathbf{a}}_l^* = 1$.

Step 4: Calculate the estimated coefficient vector based on the original data $\mathbf{D}$ as $\widehat{\mathbf{a}}_l = [\mathbf{u}_l\ \mathbf{u}_{l+1} \cdots\ \mathbf{u}_p]\ \widehat{\mathbf{a}}_l^* = \widehat{a}_{1l}^*\mathbf{u}_l + \widehat{a}_{2l}^*\mathbf{u}_{(l+1)} + \cdots \widehat{a}_{(p-1)l}^*\mathbf{u}_p = \mathbf{U}^*\widehat{\mathbf{a}}_l^*$. Return to *Step* 1.

– The orthogonal constraints are satisfied, since (for $l > 1$)

$$\begin{aligned}\widehat{\mathbf{a}}_{(l-1)}^\top\widehat{\mathbf{a}}_l = \widehat{\mathbf{a}}_{(l-1)}^\top\mathbf{U}^*\widehat{\mathbf{a}}_l^* &= \widehat{a}_{1l}^*\widehat{\mathbf{a}}_{(l-1)}^\top\mathbf{u}_l + \widehat{a}_{2l}^*\widehat{\mathbf{a}}_{(l-1)}^\top\mathbf{u}_{l+1} + \cdots + \widehat{a}_{(p-1)l}^*\widehat{\mathbf{a}}_{(l-1)}^\top\mathbf{u}_p\\ &= \widehat{a}_{1l}^*\mathbf{u}_{(l-1)}^\top\mathbf{u}_l + \widehat{a}_{2l}^*\mathbf{u}_{(l-1)}^\top\mathbf{u}_{l+1} + \cdots + \widehat{a}_{(p-1)l}^*\mathbf{u}_{(l-1)}^\top\mathbf{u}_p = 0.\end{aligned}$$

Note that, $\mathbf{u}_{(l-1)}$ and $\mathbf{u}_l$ are left singular vectors of $\mathbf{U}$ and thus, $\mathbf{u}_i^\top\mathbf{u}_j = 0$ for all $i \neq j$. Also, the orthonormal constraints are satisfied as $\widehat{\mathbf{a}}_l^T\widehat{\mathbf{a}}_l = \widehat{\mathbf{a}}_l^{*\top}\mathbf{U}^{*\top}\mathbf{U}^*\widehat{\mathbf{a}}_l^* = \widehat{\mathbf{a}}_l^{*\top}\widehat{\mathbf{a}}_l^* = 1$, since $\mathbf{U}^{*\top}\mathbf{U}^* = \mathbf{I}$.

Both algorithms were used in the simulation studies of Section 5 with generally comparable performance.

3 Influence Function

While the heuristic arguments offered point toward the inherent robustness of the index in (2), a more formal study of the robustness can be implemented through the Influence Function (IF); see Hampel et al [16], and Staudte and Sheather [30]. For a fixed $\alpha \in (0, 1)$ and dimension $k \leq p$, the IF measures the local robustness of the Rényi divergence based estimated basis $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \cdots \widehat{\mathbf{a}}_k]$ for $\mathcal{S}(\mathbf{A})$ against outlying observations. As in Section 2.6, the following derivations assume the orthonormal constraints are satisfied, but the notation Y and $\mathbf{X}$ is maintained for constancy.

For a fixed $\alpha \in (0, 1)$, let F denote the cumulative distribution function of $(Y, \mathbf{A}^\top\mathbf{X})$, then the maximization problem in (3) can be considered in terms of the functional T defined as

$$T(F) = \arg\max \mathcal{R}_\alpha(\mathbf{A}^*) = \mathbf{A}, \tag{6}$$

where $\mathbf{A}$ is a $p \times k$ matrix. To derive the influence function, let $\mathbf{W} = (Y, \mathbf{X})$ and define the contamination distribution $F_\varepsilon = (1-\varepsilon)F + \varepsilon\Delta_{\mathbf{w}_0}$, $0 < \varepsilon < 1$, where $\Delta_{\mathbf{w}_0}$ is the Dirac distribution that puts all of its mass at the point $\mathbf{w}_0 = (y_0, \mathbf{x}_0)$ and thereby, allows for the contamination of both the response and predictor vector. The influence function for T evaluated at F in the direction $\mathbf{w}_0$ is then defined as

$$\mathrm{IF}(T, F; \mathbf{w}_0) = \lim_{\varepsilon \downarrow 0} \frac{T(F_\varepsilon) - T(F)}{\varepsilon} = \frac{\partial}{\partial\varepsilon} T(F_\varepsilon)\Big|_{\varepsilon=0}, \tag{7}$$

and describes the effect of an infinitesimal amount of contamination at $\mathbf{w}_0$ on the functional T. The theoretical influence function for the index in (2) is stated in the following proposition.

Proposition 3: Let $\mathbf{A}$ denote a $p \times k$ matrix, with $k \leq p$, and in (2) define $S_\alpha(\mathbf{A}; \mathbf{w}) = \left[f(y, \mathbf{A}^\top\mathbf{x})/\{f(y)f(\mathbf{A}^\top\mathbf{x})\}\right]^{\alpha-1}$, and $\dot{S}_\alpha(\mathbf{A}; \mathbf{w}) = \frac{\partial}{\partial\mathbf{A}} S_\alpha(\mathbf{A}; \mathbf{w})$. Then, for $\alpha \in (0, 1)$, the influence function for $\mathcal{R}_\alpha(\mathbf{A})$ is given by,

$$\begin{aligned} &\mathrm{IF}(T_1, F; \mathbf{w}_0) \\ &= -\left[\left\{\int_y \int_{T(F_\varepsilon)^\top\mathbf{x}} \frac{\partial}{\partial\varepsilon}\dot{S}_\alpha(T(F_\varepsilon); \mathbf{w})\ f(y, T(F_\varepsilon)^\top\mathbf{x}) d(T(F_\varepsilon)^\top\mathbf{x}) dy\right\}\Big|_{\varepsilon=0}\right]^{-1} \dot{S}(T(F_\varepsilon); \mathbf{w}_0). \end{aligned}$$

The proof of Proposition 3 is given in Appendix A.4.

Importantly, the Rènyi based method for the robust recovery of $\mathcal{S}_{Y|\mathbf{X}}$ assumes that F is unknown and therefore, the IF needs to be estimated. The empirical distribution function $\widehat{F}$ based on a random sample $\{\mathbf{w}_i = (y_i, \mathbf{x}_i),\ i = 1, \cdots, n\}$ from $\mathbf{W} = (Y, \mathbf{X})$ can be considered by noting that based on the sample estimates in (4) and (5) of Section 2.3, the empirically based functional can be defined as $T(\widehat{F}) = \arg\max \widehat{\mathcal{R}}_\alpha(\mathbf{A}^*)$. Then, as in Critchley [10], the Empirical Sample Influence Function (ESIF) for T evaluated at $\widehat{F}$ in the direction of the i^{th} observation $\mathbf{w}_i$ is defined as

$$\mathrm{ESIF}(T, \widehat{F}, \mathbf{w}_i) = (n-1)\{T(\widehat{F}) - T(\widehat{F}_{(i)})\}, \tag{8}$$

where $\widehat{F}_{(i)} = \{1 + (n-1)^{-1}\}\widehat{F} - (n-1)^{-1}\Delta_{\mathbf{w}_i}$ is the empirical distribution function with the i^{th} observation removed.

The ESIF in (8) quantifies the influence of each observation through the change in the estimated basis when the observation is removed, which can be equivalently conceptualized as measuring the difference between the estimated subspaces of these directions; taking this into consideration Prendergast [24] suggested another Sample Influence Function (SIF) defined as

$$\text{SIF}\big(\rho_{BC}, \widehat{F}, \mathbf{w}_i\big) = (n-1)\big\{\rho_{BC}(\widehat{\mathbf{A}}_{(i)}, \widehat{\mathbf{A}}) - 1\big\}, \tag{9}$$

where $\widehat{\mathbf{A}} = T(\widehat{F})$ and $\widehat{\mathbf{A}}_{(i)} = T(\widehat{F}_{(i)})$. The ρ_{BC} term is the Bénasséni [2] Coefficient (BC) defined as

$$\rho_{BC}(\widehat{\mathbf{A}}_{(i)}, \widehat{\mathbf{A}}) = 1 - \frac{1}{k}\sum_{l=1}^{k} \big|\big|\widehat{\mathbf{a}}_l - P_{\mathcal{S}(\widehat{\mathbf{A}}_{(i)})}\,\widehat{\mathbf{a}}_l\big|\big|_2 = 1 - \frac{1}{k}\sum_{l=1}^{k} \big|\big|\big\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_{(i)})}\big\}\,\widehat{\mathbf{a}}_l\big|\big|_2, \tag{10}$$

where $\widehat{\mathbf{A}} = [\widehat{\mathbf{a}}_1\ \widehat{\mathbf{a}}_2 \cdots \widehat{\mathbf{a}}_k]$, $||\cdot||_2$ is the standard matrix 2-norm, and $P_{\mathcal{S}(\widehat{\mathbf{A}}_{(i)})} = \widehat{\mathbf{A}}_{(i)}\widehat{\mathbf{A}}_{(i)}^{\top}$ is the unique orthogonal projection matrix onto $\mathcal{S}(\widehat{\mathbf{A}}_{(i)})$; by Proposition 2, and the discussion in Section 2.6, it is assumed that $\widehat{\mathbf{A}}$ and $\widehat{\mathbf{A}}_{(i)}$ are orthonormal bases for $\mathcal{S}(\widehat{\mathbf{A}})$ and $\mathcal{S}(\widehat{\mathbf{A}}_{(i)})$, respectively. Note that, when $\mathcal{S}(\widehat{\mathbf{A}}) = \mathcal{S}(\widehat{\mathbf{A}}_{(i)})$, then $\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_{(i)})}\}$ projects $\widehat{\mathbf{a}}_l$ onto $\mathcal{S}^{\perp}(\widehat{\mathbf{A}}_{(i)})$ and consequently, $\rho_{BC}(\widehat{\mathbf{A}}_{(i)}, \widehat{\mathbf{A}}) = 1$ since $\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_{(i)})}\}\ \widehat{\mathbf{a}}_l = 0$ for all $l = 1, 2, \ldots, k$. Analogously, $\rho_{BC}(\widehat{\mathbf{A}}_{(i)}, \widehat{\mathbf{A}}) = 1$ when $\mathcal{S}(\widehat{\mathbf{A}}) = \mathcal{S}^{\perp}(\widehat{\mathbf{A}}_{(i)})$.

Due to the sensitivity of the subspace difference measure in (10) to small perturbations, it is also used as a distance measure in the bootstrap dimension detection procedure in Section 4.2. Notably, the SIF values are used directly to provide new methods for identifying the structural dimension of $\mathcal{S}_{Y|\mathbf{X}}$ and the level value of α in Sections 4.3 and 4.4, respectively.

4 Dimension estimation and α selection

4.1 Introduction

As in Section 2.6, due to the invariance of (2), the methods developed in this section assume that the orthonormal constraint $\mathbf{U}^{\top}\mathbf{U} = \mathbf{I}$ is satisfied for any matrix $\mathbf{U}$. Also, the notation

$\mathbf{U}_k$ is used with the subscript denoting the column dimension of $\mathbf{U}$.

In Section 4.2, $|\mathbf{U}|$ denotes the determinant of $\mathbf{U}$, and $||\mathbf{U}||_2$ the matrix 2-norm equal to the largest singular value of $\mathbf{U}$. Since $\mathbf{U}$ is orthonormal, $P_{\mathcal{S}(\mathbf{U})} = \mathbf{U}(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top = \mathbf{U}\mathbf{U}^\top$ is a unique orthogonal projection matrix onto $\mathcal{S}(\mathbf{U})$ and thus, $||P_{\mathcal{S}(\mathbf{U})}||_2 \leq 1$.

Note that the methods for dimension detection and tuning parameter selection described in Sections 4.3 and 4.4 are based on the SIF defined in equation (9). Consequently, these methods do not depend on the procedure used to estimate the CS, which makes the SIF methodology applicable more generally. It is argued below that the approach based on the SIF values is intuitively more sensible than bootstrap methodologies in the presence of contamination, which is supported by the results of the simulation studies presented in Web Appendix C using the Rényi divergence based method. An interesting future direction would be to investigate whether the improved performance of the SIF based methods over bootstrapping approaches continues to hold for other existing SDR methods.

4.2 Bootstrap dimension estimation

For a fixed $\alpha \in (0, 1)$, let $\mathbf{X}$ denote a $p\times 1$ random predictor vector and $\mathbf{A}_k$ a k-dimensional basis for a DRS for the regression of Y on $\mathbf{X}$, where $k \in \{1, 2, \ldots, (p-1)\}$. In general, the structural dimension for the regression of Y on $\mathbf{X}$ can be considered as the value of k such that the subspace spanned by the columns of $\mathbf{A}_k$, $\mathcal{S}(\mathbf{A}_k)$, has the least variability. More specifically, the dimension of $\mathcal{S}_{Y|\mathbf{X}}$ can be identified as the value $k \in \{1, 2, \ldots, (p-1)\}$ that produces the smallest averaged bootstrapped estimate of the variability of $\mathcal{S}(\mathbf{A}_k)$. The dimension $k < p$ is assumed since $k = p$ leads to the trivial case $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{I}_p^\top\mathbf{X}$; the effect on the distance measures used in the bootstrap procedure when $k = p$ is discussed below.

First, to estimate the variability of $\mathcal{S}(\mathbf{A}_k)$ for each fixed k, an estimate $\widehat{\mathbf{A}}_k$ of $\mathbf{A}_k$ is calculated based on the observed dataset $\mathbf{D} = \{(y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n)\}$. Next, a bootstrap estimate $\widehat{\mathbf{A}}_k^b$ of $\mathbf{A}_k^b$ is calculated from a bootstrapped dataset $\mathbf{D}^b = \{(y_1, \mathbf{x}_1)^b, (y_2, \mathbf{x}_2)^b, \ldots, (y_n, \mathbf{x}_n)^b\}$ that is generated by randomly sampling from $\mathbf{D}$ with replacement. For each of the $b = 1, 2, \ldots, B$ bootstrap iterations, a distance between the subspaces $\mathcal{S}(\widehat{\mathbf{A}}_k)$ and $\mathcal{S}(\widehat{\mathbf{A}}_k^b)$ is calculated and averaged over all iterations to give an estimate of the variability of $\mathcal{S}(\mathbf{A}_k)$. This method has been well studied in a multivariate association context in the papers of Iaci et al [22] and Ye and Weiss [36], where they used many different subspace distance measures. Three different distances are investigated, including a new estimate based on the Bénasséni coefficient (BC), all defined next.

To measure the distance between $\mathcal{S}(\widehat{\mathbf{A}}_k)$ and $\mathcal{S}(\widehat{\mathbf{A}}_k^b)$ consider a slight modification of the BC given in (10) and define

$$\rho_{_{BC^*}}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k) = \frac{1}{k}\sum_{l=1}^{k} \left|\left|\widehat{\mathbf{a}}_l - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\, \widehat{\mathbf{a}}_l\right|\right|_2 = \frac{1}{k}\sum_{l=1}^{k} \left|\left|\left\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right\} \widehat{\mathbf{a}}_l\right|\right|_2, \tag{11}$$

where $P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)} = \widehat{\mathbf{A}}_k^b \widehat{\mathbf{A}}_k^{b^\top}$ is the unique projection matrix onto $\mathcal{S}(\widehat{\mathbf{A}}_k^b)$. When $\mathcal{S}(\widehat{\mathbf{A}}_k) = \mathcal{S}(\widehat{\mathbf{A}}_k^b)$, then $\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\}$ projects $\widehat{\mathbf{a}}_l$ onto $\mathcal{S}^\perp(\widehat{\mathbf{A}}_k)$ and consequently, $\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\}\ \widehat{\mathbf{a}}_l = 0$ for all $l = 1, 2, \ldots, k$ and therefore, $\rho_{_{BC^*}}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k) = 0$. Also, $\rho_{_{BC^*}}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k) \leq 1$ since $\left|\left|\left\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right\} \widehat{\mathbf{a}}_l\right|\right|_2 \leq \left|\left|\left\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right\}\right|\right|_2 \left|\left|\widehat{\mathbf{a}}_l\right|\right|_2 = \left|\left|\left\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right\}\right|\right|_2 \leq 1$; the bound a result of the discussion in Section 4.1. Therefore, smaller values of (11) correspond to more equivalent subspaces.

Next, the L_2 norm subspace distance investigated in Iaci et. al [22] is used and defined as

$$L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k) = \left|\left|P_{\mathcal{S}(\widehat{\mathbf{A}}_k)}\left\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right\}\right|\right|_2. \tag{12}$$

Note that, if $\mathcal{S}(\widehat{\mathbf{A}}_k) = \mathcal{S}(\widehat{\mathbf{A}}_k^b)$, then $\left(\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\right)$ projects onto $\mathcal{S}^\perp(\widehat{\mathbf{A}}_k)$ and $L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k) = 0$, and $\left|\left|P_{\mathcal{S}(\widehat{\mathbf{A}}_k)}\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\}\right|\right|_2 \leq \left|\left|P_{\mathcal{S}(\widehat{\mathbf{A}}_k)}\right|\right|_2 \left|\left|\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\}\right|\right|_2 \leq 1$. Again, small values of (12) indicate similar subspaces.

The last distance investigated is the one used in Ye and Weiss [36], and then in Iaci et. al [22], based on the square root of Hotelling's [19] squared vector correlation coefficient, given by

$$1 - \rho_{_{HC}}\left(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\right) = 1 - \left|\widehat{\mathbf{A}}_k^\top \widehat{\mathbf{A}}_k^b \widehat{\mathbf{A}}_k^{b^\top} \widehat{\mathbf{A}}_k\right|^{\frac{1}{2}} = 1 - \left(\prod_{i=1}^{\mathrm{k}} \lambda_i\right)^{\frac{1}{2}}, \tag{13}$$

where the $\lambda_i, i = 1, 2, \ldots, k$, are the eigenvalues of $\widehat{\mathbf{A}}_k^\top \widehat{\mathbf{A}}_k^b \widehat{\mathbf{A}}_k^{b^\top} \widehat{\mathbf{A}}_k$. Since $\rho_{_{HC}}\left(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\right)$ is a measure of the correlation between $\mathcal{S}(\widehat{\mathbf{A}}_k)$ and $\mathcal{S}(\widehat{\mathbf{A}}_k^b)$, with $\rho_{_{HC}}\left(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\right) = 1$ when the subspaces are equal, and $\rho_{_{HC}}\big(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\big) = 0$ when they are orthogonal, smaller values of $1 - \rho_{_{HC}}\big(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\big)$ correspond to similar subspaces.

For a fixed k, small values of (11), (12) and (13) correspond to similarly estimated subspaces based on the original and bootstrapped datasets, where both are meant to estimate $\mathcal{S}(\mathbf{A})$. Therefore, small values of $\overline{\rho}_{_{BC^*}} = \frac{1}{B}\sum_{b=1}^{B} \rho_{_{BC^*}}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$, $\overline{L}_{2(O)} = \frac{1}{B}\sum_{b=1}^{B} L_{2(O)}(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k)$, or $1 - \overline{\rho}_{_{HC}} = 1 - \frac{1}{B}\sum_{b=1}^{B} \rho_{_{HC}}\big(\widehat{\mathbf{A}}_k^b, \widehat{\mathbf{A}}_k\big)$, $b = 1, 2, \ldots, B$, identify a dimension k where $\mathcal{S}(\mathbf{A}_k)$ has less variability. The structural dimension d can then be estimated using any of these three methods by determining the value $k^* \in \{1, 2, \ldots, (p-1)\}$ that produces the smallest value, and setting $\widehat{d} = k^*$. Alternatively, for each $k < p$, boxplots of the individual values of (11), (12) or (13) for each of the bootstrap iterations can be created to visually determine the value k^* that corresponds to the boxplot that is the most closely centered near zero with the smallest variability.

Note that when $k = p$, all measures are zero, since $\widehat{\mathbf{A}}_k$ and $\widehat{\mathbf{A}}_k^b$ span the same vector space $\mathbf{R}^p$. Specifically, as discussed in Section 4.1, under the orthonormal constraints the orthogonal projections $P_{\mathcal{S}(\widehat{\mathbf{A}}_k)}$ and $P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}$ onto $\mathbf{R}^p$ are unique and thus, $\big(\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\big) = \mathbf{0}$ and $|\widehat{\mathbf{A}}_k^\top \widehat{\mathbf{A}}_k^b \widehat{\mathbf{A}}_k^{b^\top} \widehat{\mathbf{A}}_k| = |\mathbf{I}|$ in (12) and (13), respectively; and for the distance measure in (11), every term in the summation is $||\big\{\mathbf{I} - P_{\mathcal{S}(\widehat{\mathbf{A}}_k^b)}\big\}\widehat{\mathbf{a}}_l||_2 = ||\{\mathbf{I}_p - \mathbf{I}_p\}\widehat{\mathbf{a}}_l||_2 = 0$.

When the dimension of the estimated coefficient matrix exceeds the true structural dimension, $k > d$, consider the partition $\widehat{\mathbf{A}}_k = [\widehat{\mathbf{A}}_d \ \widehat{\mathbf{A}}_{(k-d)}]$, where $\widehat{\mathbf{A}}_d$ is the estimated basis of $\mathcal{S}_{Y|\mathbf{X}}$ and $\widehat{\mathbf{A}}_{(k-d)}$ an estimated basis of a (k - d) dimensional subspace orthogonal to $\mathcal{S}_{Y|\mathbf{X}}$. Then, the resulting estimates of the partial coefficient matrices $\widehat{\mathbf{A}}_{(k-d)}$, and analogously $\widehat{\mathbf{A}}_{(k-d)}^b$, will be determined randomly in subspaces orthogonal to $\mathcal{S}(\widehat{\mathbf{A}}_d)$ and $\mathcal{S}(\widehat{\mathbf{A}}_d^b)$, respectively. In general, for all three distance measures, this will be reflected through larger values of the distance between $\mathcal{S}(\widehat{\mathbf{A}}_k)$ and $\mathcal{S}(\widehat{\mathbf{A}}_k^b)$. However, different from the distance measures in (12) and (13), $\rho_{_{BC^*}}$ averages over each dimension $k < p$ and thus, for $d << p$ it trends to 0 when $k >> d$ in practice. In such a case, the centers of the boxplots, or bar plots of the mean, of $\rho_{_{BC^*}}$ can show a decreasing parabolic pattern as k approaches p. For example, in Simulation I of Study 1 the predictor vector has dimension $p = 10$ and the structural dimension is $d = 1$. The top left panel of Web Appendix C.3

Figure 3 shows that the $\rho_{_{BC^*}}$ values begin to decrease for $k > 3$. However, it is clear that there is a significant increase in the centers of the boxplots from $k = 1$ to $k = 2, 3$ and therefore, the estimated true dimension is taken to be $\widehat{d} = 1$. That is, the centers of the boxplots for $d < k << p$ will make the estimated $\widehat{d} = k^* \in \{1, 2, \ldots, (p-1)\}$ discernible.

4.3 SIF dimension estimation

Bootstrapping a dataset that contains outlying observations can increase the level of contamination, making the estimated subspaces spanned by $\widehat{\mathbf{A}}_k$ and $\widehat{\mathbf{A}}_k^b$ more variable even when $k = d$. For example, this can be seen in the boxplots in the bottom right panel of Web Appendix C.4 Figure 4 of the bootstrapped values of the $L_{2(O)}$ distance in (12) for Simulation I of Study 2. For this simulation the true structural dimension is $d = 2$ and the level of contamination is 10%, and assuming an exact number of 30 contaminated values based on a sample size of $n = 300$, the probability of maintaining or increasing the level of contamination of each bootstrapped dataset is 0.52. Further, the spirit of the bootstrap procedure is that for $k > d$, the distance and variability between the estimated and bootstrap estimated subspaces increases, which is comparable to all data values becoming influential. For example, this can be seen in the smoothed curve plots of the scaled SIF values for Simulation I of Study 2 in the far right panel of Figure 2, where the plots corresponding to the dimensions $k = 3$ and 4 are all well below the SIF curves corresponding to $k = 1$ and $d = k^* = 2$ and thus, show that most of the $n = 300$ observations are much more influential at these dimensions. This motivates the investigation of a new method for estimating d based on the SIF values.

For a fixed α the true dimension d can be estimated more efficiently using the sample influence values since they are a direct measure of the effect of the observations on the estimated subspace $\mathcal{S}(\widehat{\mathbf{A}}_k)$. That is, the estimated structural dimension can be defined as the dimension that produces the fewest influential observations among the possible dimensions $k < p$. Specifically, the absolute value of (9), $|s_{(i,k)}| = \left|\mathrm{SIF}\left(\rho_{_{BC}}, \widehat{F}, \mathbf{w}_i\right)\right| = \left|(n-1)\left\{\rho_{_{BC}}(\widehat{\mathbf{A}}_{(i)k}, \widehat{\mathbf{A}}_k) - 1\right\}\right|$, is calculated for each observation $\mathbf{w}_i = (y_i, \mathbf{x}_i), i = 1, 2, \ldots, n$, at each dimension $k \in \{1, 2, \ldots, (p-1)\}$. Note that, the absolute value of the SIF measures are taken to be comparable to the bootstrap plots, and $\widehat{\mathbf{A}}_{(i)k}$ is the estimated coefficient matrix with the i^{th} observation removed. Next, as in the above bootstrap procedure, boxplots of the $|s_{(i,k)}|$ values, $i = 1, 2, \ldots, n$, or bar plots of the means, for each k can be generated to visually determine the estimated true dimension $\widehat{d}$. Specifically, the dimension k^* such that the centers and spread increase noticeably for $k > k^*$ is taken to be the estimated structural dimension $\widehat{d} = k^*$. Also, smoothed values of the raw SIF values can be plotted to determine the first dimension k^* such that for $k > k^*$ a significantly larger amount of observations are highly influential. All three visualizations methods are used in the numerical studies in Section 5, and in practice all can be used to confirm the inference made for the dimension of $\mathcal{S}_{Y|\mathbf{X}}$.

4.4 SIF tuning parameter selection

For a fixed dimension k, the SIF values can also provide a powerful and efficient method for determining the level of the tuning parameter that generates the most robust estimate of $\mathbf{A}_k$. Intuitively, the value of $\alpha \in (0, 1)$ that parameterizes the most robust index in (4) is the value that produces an estimate of $\mathbf{A}_k$ that is the least sensitive to influential observations. Let $H = \{\alpha_1, \alpha_2, \ldots, \alpha_m\}$ denote a set of m different values of the tuning parameter, then quantifying robustness at different values of α is naturally accomplished by first considering the SIF values in (9), $s_\alpha(\mathbf{w}_i) = \text{SIF}\big(\rho_{BC}, \widehat{F}, \mathbf{w}_i\big) = (n-1)\big\{\rho_{BC}(\widehat{\mathbf{A}}_{(i)k}, \widehat{\mathbf{A}}_k) - 1\big\}$, for each observation $\mathbf{w}_i = (y_i, \mathbf{x}_i), i = 1, 2, \ldots, n$, at a fixed level of $\alpha \in H$. Next, a smoothed plot of the ordered $s_\alpha(\mathbf{w}_i), i = 1, 2, \ldots, n$, can be used to visually determine the level $\alpha \in H$ that *dominates* all other values of the tuning parameter. For example, for Simulation I of Study 2 the top left panel of Web Appendix C.4 Figure 4 shows the plots of the ordered SIF values at four different levels of α, with $\alpha = 0.8$ generally producing the smallest values of $s_\alpha(\mathbf{w}_i)$ for all observations. In this sense, $\alpha = 0.8$ dominates all other levels considered.

Therefore, the optimal level of α for robustness can be identified by considering the area above each of the plots of the smoothed ordered SIF values as shown in the left panel of Figure 2. Equivalently, this can be measured as the area under the curve (AUC) of the smoothed plot of the $|s_\alpha(\mathbf{w}_i)|$ values, as shown in the middle panel of Figure 2. To quantify this specifically, for a fixed α the AUC is calculated on the absolute value of the ascending ordered SIF values using the trapezoidal rule as $\text{AUC}_\alpha = \frac{1}{2}\sum_{h=1}^{n-1}(u_{h+1} - u_h)\{|s_\alpha(\mathbf{w}_h)| + |s_{\alpha,h+1}(\mathbf{w}_{h+1})|\} = \frac{1}{2}\sum_{h=1}^{n-1}\{|s_{\alpha,h}(\mathbf{w}_h)|+|s_{\alpha,h+1}(\mathbf{w}_{h+1})|\}$, where $\{u_h = h; h = 1, 2, \dots, (n-1)\}$ is taken to create subintervals of unit length. A simple comparison of the AUC_α measures for different values of α can then be used to select the value that parameterizes the most robust sample index. For example, in the previously mentioned simulation the top right panel of Web Appendix C.4 Figure 4 shows that $\alpha = 0.8$ produces the lowest AUC among the nine different values selected for comparison.

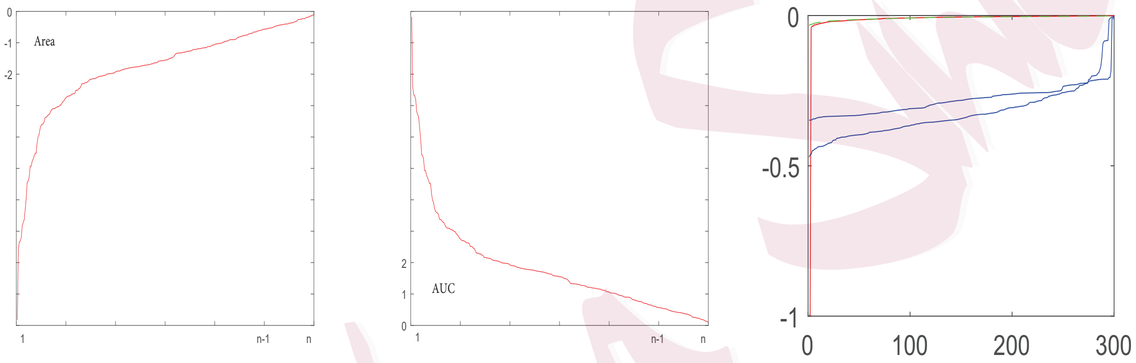

Figure 2: Left panel: example plot of $s_\alpha(\mathbf{w}_i)$ with labeled Area; Middle panel: example plot of $|s_\alpha(\mathbf{w}_i)|$ with AUC_α labeled AUC; Right panel: simulation plot of $s_\alpha(\mathbf{w}_i)$ for dimensions $k = 1, 2, 3, 4$ with $d = 2$.

5 Simulation studies

In this section, we introduce the various regression models and parameters of the three different numerical studies used to investigate the performance of our method in estimating a basis for $\mathcal{S}_{Y|\mathbf{X}}$ in the presence of data contamination. A detailed, self-contained, discussion of the simulation studies and results are provided in Web Appendix C.

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 $\mathbf{X} = (X_1, \dots, X_{10})^\top$ 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 [24] 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.

As suggested by a reviewer, an additional simulation study is implemented to compare the performance of our method to other robust methods for estimating a basis of the CS. To this end, the performance of the Rényi based method is compared to the results reported in Zhang [42] under the same simulation parameters. A motivation for considering their study is that they demonstrated an improved performance using their robust methods for estimating a basis of $\mathcal{S}_{Y|\mathbf{X}}$, termed regMAVE and regOPG, over SIR, SAVE, MAVE, and the robust MAVE (rtMAVE) method of Čížek and Härdle [5], for the regression model of Study 3 in the presence of symmetrically contaminated error terms generated from a $U(-\theta, \theta)$ distribution. Note that, Zhang et al. [42] defined the error term in the regression model as $0.5\varepsilon^*$, where $\varepsilon^* \sim U(-\theta, \theta)$, which would be equivalent to generating contaminated error terms from a $U(-0.5\theta, 0.5\theta)$ distribution for the error term defined in the simulation studies above. Also, different from Simulation Study 3, Zhang et al. [42] investigated a multivariate normal explanatory vector with a Toeplitz matrix covariance dependence structure. The full details and results of this numerical study are provided in Web Appendix E, where it is shown that our Rényi divergence based approach estimates the CS with improved accuracy.

6 Baseball salary data

The inherent robustness of our method is illustrated in the analysis of a well-studied dataset, known to contain outliers and extremes observations, that was initially presented in a sponsored section on statistics and graphics of the American Statistical Association in 1998 with the stated goal of answering the question, “are players paid according to their performance?” Importantly, different from many previous 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. A more comprehensive discussion of the dataset with a detailed analysis is given in Web Appendix D.

7 Supplementary material

The supplementary material is prepared as an extension of Appendix A and for this reason is referenced in the main text as the Web Appendix. The first section, Web Appendix B, provides the full heuristic argument for robustness mentioned in Section 2.4. The simulation methodology of Section 5 is comprehensively discussed in Web Appendix C, with measures to quantify the accuracy of the estimated central subspaces given in Web Appendix C.1, regression models and parameters provided in Web Appendix C.2, and the results reported in Web Appendices C.3-C.5. As a demonstration of our methodology in a real world application, a complete discussion and analysis of the baseball salary dataset introduced in Section 6 is provided in Web Appendix D.

Acknowledgment

We would like to dedicate this work to our late colleague Professor Xiangrong Yin. We would like to thank the Editor, an Associate Editor, and a reviewer for their insightful comments that led to an improvement of this paper. Iaci's work was supported by NSF grant 1309954 and Sriram's work supported by NSF grant 1205546.

A Appendix

A.1 Proof of Proposition 1

Consider the alternate expression of $\mathcal{R}_\alpha(\mathbf{A})$,

$$\begin{aligned}\mathcal{R}_\alpha(\mathbf{A}) = \frac{1}{\alpha-1}\ln\left(\mathrm{E}\left[\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^\alpha\left\{\frac{f(Y)f(\mathbf{A}^\top\mathbf{X})}{f(Y,\mathbf{A}^\top\mathbf{X})}\right\}\right]\right)\\ = \frac{1}{\alpha-1}\ln\left[\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left\{\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{A}^\top\mathbf{x})f(y)\,d(\mathbf{A}^\top\mathbf{x})\,dy\right]\\ = \frac{1}{\alpha-1}\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^\alpha\right], \qquad (14)\end{aligned}$$

and note the following result.

Result 1: For $\alpha \in (0,1)$, define $h(l) = l^\alpha, l > 0$, then $h^{''}(l) = \alpha(\alpha-1)l^{\alpha-2} < 0 \implies h(l)$ is concave for all l. Then, for the random variable L, $E\{h(L)\} \leq h\{E(L)\}$ by Jensen's inequality.

Suppose that $\mathcal{S}(\mathbf{A}_1) \subseteq \mathcal{S}(\mathbf{A})$, then $\mathbf{A}_1 = \mathbf{AB}$ for some matrix $\mathbf{B}$ such that $rank(\mathbf{A}_1) \equiv rank(\mathbf{AB}) \leq rank(\mathbf{A}) \implies f(y|\ \mathbf{B}^\top\mathbf{w}, \mathbf{A}^\top\mathbf{x}) = f(y|\ \mathbf{A}^\top\mathbf{x})$, where $\mathbf{w} = \mathbf{A}^\top\mathbf{x}$ and $\mathbf{A}_1^\top\mathbf{x} = \mathbf{B}^\top\mathbf{A}^\top\mathbf{x} = \mathbf{B}^\top\mathbf{w}$. This leads to the following result.

Result 2:

$$f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x},y)=\frac{f(y,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y,\mathbf{A}^\top\mathbf{x})}=\frac{f(y|\ \mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})f(\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\ \mathbf{A}^\top\mathbf{x})f(\mathbf{A}^\top\mathbf{x})}$$
$$=\frac{f(y|\ \mathbf{A}^\top\mathbf{x})f(\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\ \mathbf{A}^\top\mathbf{x})f(\mathbf{A}^\top\mathbf{x})}$$
$$=f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x}),$$

which implies $f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x},y)/f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x})=1$.

For the alternate expression in (14), this leads to the following result.

Result 3:

$$\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left\{\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{A}^\top\mathbf{x})f(y)\ d(\mathbf{A}^\top\mathbf{x})dy$$
$$=\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x},y)}{f(\mathbf{B}^\top\mathbf{w}|\ \mathbf{A}^\top\mathbf{x})}\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})\ d(\mathbf{B}^\top\mathbf{w})\right]f(y)\ d(\mathbf{A}^\top\mathbf{x})dy$$
$$=\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(y,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})f(\mathbf{B}^\top\mathbf{w})\ d(\mathbf{B}^\top\mathbf{w})\right]f(y)\ d(\mathbf{A}^\top\mathbf{x})dy$$
$$=\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left[\int_{\mathbf{A}^\top\mathbf{x}}\left\{\frac{f(y,\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})}{f(y|\ \mathbf{B}^\top\mathbf{w})f(\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})}\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}\right\}^\alpha f(\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})\ d(\mathbf{A}^\top\mathbf{x})\right]$$
$$\times f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy$$
$$\leq\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left[\int_{\mathbf{A}^\top\mathbf{x}}\frac{f(y,\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})}{f(y|\ \mathbf{B}^\top\mathbf{w})f(\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})}\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}f(\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})\ d(\mathbf{A}^\top\mathbf{x})\right]^\alpha$$
$$\times f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy$$
$$=\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}\right\}^\alpha\left[\int_{\mathbf{A}^\top\mathbf{x}}\frac{f(y,\mathbf{A}^\top\mathbf{x}|\ \mathbf{B}^\top\mathbf{w})}{f(y|\ \mathbf{B}^\top\mathbf{w})}\ d(\mathbf{A}^\top\mathbf{x})\right]^\alpha f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy$$
$$=\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}\right\}^\alpha\left[\int_{\mathbf{A}^\top\mathbf{x}}\frac{f(y,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\ \mathbf{B}^\top\mathbf{w})f(\mathbf{B}^\top\mathbf{w})}\ d(\mathbf{A}^\top\mathbf{x})\right]^\alpha f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy$$
$$=\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}\right\}^\alpha\left[\int_{\mathbf{A}^\top\mathbf{x}}f(\mathbf{A}^\top\mathbf{x}|\ y,\mathbf{B}^\top\mathbf{w})\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y,\mathbf{B}^\top\mathbf{w})}\ d(\mathbf{A}^\top\mathbf{x})\right]^\alpha$$
$$\times f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy$$
$$=\int_y\int_{\mathbf{B}^\top\mathbf{w}}\left\{\frac{f(y,\mathbf{B}^\top\mathbf{w})}{f(y)f(\mathbf{B}^\top\mathbf{w})}\right\}^\alpha f(\mathbf{B}^\top\mathbf{w})f(y)\ d(\mathbf{B}^\top\mathbf{w})dy,$$

where the first equality stems from Result 2 and the inequality due to Result 1. Therefore,

$$\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^\alpha\right]\leq\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{B}^\top\mathbf{W})}{f(Y)f(\mathbf{B}^\top\mathbf{W})}\right\}^\alpha\right]=\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{A}_1^\top\mathbf{X})}{f(Y)f(\mathbf{A}_1^\top\mathbf{X})}\right\}^\alpha\right]$$
$$\implies\frac{1}{\alpha-1}\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^\alpha\right]\geq\frac{1}{\alpha-1}\ln\left[\mathrm{E}^*\left\{\frac{f(Y,\mathbf{A}_1^\top\mathbf{X})}{f(Y)f(\mathbf{A}_1^\top\mathbf{X})}\right\}^\alpha\right],\ \text{since }(\alpha-1)^{-1}<0,$$

which shows that $\mathcal{R}_\alpha(\mathbf{A})\geq\mathcal{R}_\alpha(\mathbf{A}_1)$ when $\mathcal{S}(\mathbf{A}_1)\subseteq\mathcal{S}(\mathbf{A})$.

Next, suppose that $\mathcal{S}(\mathbf{A}_1)=\mathcal{S}(\mathbf{A})$, then $\mathbf{A}_1=\mathbf{AB}$ for some matrix $\mathbf{B}$ such that $rank(\mathbf{A}_1)\equiv rank(\mathbf{AB})=rank(\mathbf{A})\implies f(y|\ \mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})=f(y|\ \mathbf{B}^\top\mathbf{x})$ for all $y,\mathbf{B}^\top\mathbf{w}$

and $\mathbf{A}^\top\mathbf{x}$. This implies that

$$\begin{aligned}\frac{f(y,\mathbf{A}^\top\mathbf{x}|\,\mathbf{B}^\top\mathbf{w})}{f(y|\,\mathbf{B}^\top\mathbf{w})f(\mathbf{A}^\top\mathbf{x}|\,\mathbf{B}^\top\mathbf{w})} = \frac{f(y,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{B}^\top\mathbf{w})f(\mathbf{A}^\top\mathbf{x}|\,\mathbf{B}^\top\mathbf{w})f(\mathbf{B}^\top\mathbf{w})} &= \frac{f(y,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{B}^\top\mathbf{w})f(\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}\\ &= \frac{f(y|\,\mathbf{B}^\top\mathbf{w},\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{B}^\top\mathbf{x})}\\ &= \frac{f(y|\,\mathbf{B}^\top\mathbf{x})}{f(y|\,\mathbf{B}^\top\mathbf{x})} = 1\end{aligned}$$

and therefore, the inequality becomes equality $\implies \mathcal{R}_\alpha(\mathbf{A}) = \mathcal{R}_\alpha(\mathbf{A}_1)$.

By definition, $\mathcal{S}(\mathbf{A}) \subseteq \mathcal{S}(\mathbf{I})$ since $k \le p \implies \mathcal{R}_\alpha(\mathbf{I}) \ge \mathcal{R}_\alpha(\mathbf{A})$ as shown above. Next, setting $\mathbf{A}^\top\mathbf{x} = \mathbf{I}^\top\mathbf{x} = \mathbf{x}$ and $\mathbf{B}^\top\mathbf{w} = \mathbf{A}^\top\mathbf{x}$, Result 2 holds and thus, substituting into Result 3 leads to

$$\begin{aligned}&\int_y\int_{\mathbf{x}}\left\{\frac{f(y,\mathbf{x})}{f(y)f(\mathbf{x})}\right\}^\alpha f(\mathbf{x})f(y)\;d\mathbf{x}dy\\ &= \int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[\int_{\mathbf{x}}\left\{\frac{f(y,\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{A}^\top\mathbf{x})f(\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})\;d\mathbf{x}\right]f(\mathbf{A}^\top\mathbf{x})f(y)\;d(\mathbf{A}^\top\mathbf{x})dy\\ &\le \int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[\int_{\mathbf{x}}\frac{f(y,\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{A}^\top\mathbf{x})f(\mathbf{A}^\top\mathbf{x})}f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})\;d\mathbf{x}\right]^\alpha f(\mathbf{A}^\top\mathbf{x})f(y)\;d(\mathbf{A}^\top\mathbf{x})dy\\ &= \int_y\int_{\mathbf{A}^\top\mathbf{x}}\left\{\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\right\}^\alpha f(\mathbf{A}^\top\mathbf{x})f(y)\;d(\mathbf{A}^\top\mathbf{x})dy.\end{aligned}$$

Next, since $h(l)$ is strictly concave for $\alpha \in (0,1)$, equality holds if and only if for all $\mathbf{y}$, $\mathbf{A}^\top\mathbf{x}$ and $\mathbf{x}$,

$$\begin{aligned}\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})} &= \frac{f(y,\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}{f(y|\,\mathbf{A}^\top\mathbf{x})f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}\frac{f(y,\mathbf{A}^\top\mathbf{x})}{f(y)f(\mathbf{A}^\top\mathbf{x})}\\ &\implies \frac{f(y,\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})}{f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})f(y|\,\mathbf{A}^\top\mathbf{x})} = 1 \implies f(y,\mathbf{x}|\,\mathbf{A}^\top\mathbf{x}) = f(y|\,\mathbf{A}^\top\mathbf{x})f(\mathbf{x}|\,\mathbf{A}^\top\mathbf{x})\end{aligned}$$

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

A.2 Proof of Proposition 2

First, note the equivalent conditional expectation representation of (2),

$$\mathrm{E}_{(Y,\mathbf{A}^\top\mathbf{X})}\{f(Y,\mathbf{A}^\top\mathbf{X})/f(Y)f(\mathbf{A}^\top\mathbf{X})\}^{\alpha-1} = \mathrm{E}_{\mathbf{A}^\top\mathbf{X}}\mathrm{E}_{(Y|\mathbf{A}^\top\mathbf{X})}\{f(Y|\,\mathbf{A}^\top\mathbf{X})/f(Y)\}^{\alpha-1},$$

since

$$\begin{aligned}&\int_y\int_{\mathbf{A}^\top\mathbf{x}}\{f(Y,\mathbf{A}^\top\mathbf{X})/f(Y)f(\mathbf{A}^\top\mathbf{X})\}^{\alpha-1}f(y,\mathbf{A}^\top\mathbf{x})\;d(\mathbf{A}^\top\mathbf{x})dy\\ &\qquad\qquad = \int_{\mathbf{A}^\top\mathbf{x}}\int_y\{f(Y|\,\mathbf{A}^\top\mathbf{X})/f(Y)\}^{\alpha-1}f(y|\,\mathbf{A}^\top\mathbf{x})f(\mathbf{A}^\top\mathbf{x})\;dy\;d(\mathbf{A}^\top\mathbf{x}).\end{aligned}$$

Also,

$$\mathrm{E}_{(Y,\mathbf{A}^\top\mathbf{X})}\{f(Y,\mathbf{A}^\top\mathbf{X})/f(Y)f(\mathbf{A}^\top\mathbf{X})\}^{\alpha-1} = \mathrm{E}_Y\mathrm{E}_{(\mathbf{A}^\top\mathbf{X}|Y)}\{f(\mathbf{A}^\top\mathbf{X}|\,Y)/f(\mathbf{A}^\top\mathbf{X})\}^{\alpha-1}$$

and thus,

$$\begin{aligned}
(\alpha-1)\mathcal{R}_{(Y,\mathbf{X})}(\mathbf{A}) &= \ln\Big[\mathrm{E}_{(Y,\mathbf{A}^\top\mathbf{X})}\{f(Y,\mathbf{A}^\top\mathbf{X})/f(Y)f(\mathbf{A}^\top\mathbf{X})\}^{\alpha-1}\Big]\\
&= \ln\Big[\mathrm{E}_{\mathbf{A}^\top\mathbf{X}}\mathrm{E}_{(Y|\mathbf{A}^\top\mathbf{X})}\{f(Y|\ \mathbf{A}^\top\mathbf{X})/f(Y)\}^{\alpha-1}\Big]\\
&= \ln\Big[\mathrm{E}_{\mathbf{A}^\top\mathbf{C}_2\{\mathbf{W}_2-\mathbf{b}\}}\mathrm{E}_{(Y|\mathbf{A}^\top\mathbf{C}_2\{\mathbf{W}_2-\mathbf{b}\})}\{f(Y|\ \mathbf{A}^\top\mathbf{C}_2\{\mathbf{W}_2-\mathbf{b}\})/f(Y)\}^{\alpha-1}\Big]\\
&= \ln\Big[\mathrm{E}_{(\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)}\mathrm{E}_{(Y|\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)}\{f(Y|\ \mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)/f(Y)\}^{\alpha-1}\Big]\\
&= \ln\Big[\mathrm{E}_{(Y,\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)}\{f(Y,\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)/f(Y)f(\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)\}^{\alpha-1}\Big] \qquad (15)
\end{aligned}$$

- conditioning on Y in (15) and using the analogous steps above -

$$\begin{aligned}
&= \ln\Big[\mathrm{E}_{(C_1W_1,\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)}\{f(C_1W_1,\mathbf{A}^\top\mathbf{C}\mathbf{W}_2)/f(C_1W_1)f(\mathbf{A}^\top\mathbf{C}_2\mathbf{W}_2)\}^{\alpha-1}\Big]\\
&= (\alpha-1)\mathcal{R}_{(W_1,\mathbf{W}_2)}(\mathbf{C}_2^\top\mathbf{A}).
\end{aligned}$$

Therefore, $\mathcal{R}_{(Y,\mathbf{X}),\alpha}(\mathbf{A}) = \mathcal{R}_{(W_1,\mathbf{W}_2),\alpha}(\mathbf{C}_2^\top\mathbf{A})$, and setting $\mathbf{A} = \mathbf{I} \implies \mathcal{R}_{(Y,\mathbf{X}),\alpha}(\mathbf{I}) = \mathcal{R}_{(W_1,\mathbf{W}_2),\alpha}(\mathbf{C}_2^\top\mathbf{I})$.

A.3 Proof of consistency

Let $\mathbf{U}_i$ be a sequence of k-dimensional random vectors with distribution function F and Lebesgue measurable density f. Define the kernel density estimate of f as

$$f_n(\mathbf{u}) = \frac{1}{na_n^k}\sum_{i=1}^{n}\mathrm{K}\left(\frac{\mathbf{u}-\mathbf{U}_i}{a_n^k}\right) \quad \text{for } \mathbf{u}\in\mathbb{R}^k,$$

where $\mathrm{K}:\mathbb{R}^k\to\mathbb{R}^+$ is a probability density on $\mathbb{R}^k$, uniformly for $\|\mathbf{u}\|\to\infty$, and $a_n>0$ such that $\lim_{n\to\infty}a_n=0$. Noting that the theorem 1-m of Kiefer [23] holds for all F, a direct application of Theorem 1 of Ruschendorf [28] yields the following lemma.

Lemma 1 *Let $\{(y_i,\mathbf{x}_i); i=1,2,\dots,n\}$ be iid and*

$$\sum_{n=1}^{\infty} e^{-\gamma n a_n^{2k_r}} < \infty \textit{ for all } \gamma > 0.$$

Let K be of bounded variation and $f(y), f(\mathbf{A}^\top\mathbf{x})$ and $f(y,\mathbf{A}^\top\mathbf{x})$ be uniformly continuous in $y, \mathbf{A}$ and $\mathbf{x}$. Under these conditions as $n\to\infty$:

$$\begin{aligned}
\sup_{y\in\mathbb{R}}|f_n(y)-f(y)| &\to 0 \ \ a.s., \ \textit{where } k_r = 1\\
\sup_{\mathbf{A},\ \mathbf{x}\in\mathbb{R}^k}\left|f_n(\mathbf{A}^\top\mathbf{x})-f(\mathbf{A}^\top\mathbf{x})\right| &\to 0 \ \ a.s., \ \textit{where } k_r = k\\
\sup_{\mathbf{A},\ y\in\mathbb{R},\ \mathbf{x}\in\mathbb{R}^k}\left|f_n\left(y,\mathbf{A}^\top\mathbf{x}\right)-f\left(y,\mathbf{A}^\top\mathbf{x}\right)\right| &\to 0 \ \ a.s., \ \textit{where } k_r = 1+k.
\end{aligned}$$

Proof of Theorem 1: Assume that the conditions of Lemma 1 hold, and let $\epsilon>0, b>0\to 0$ as $n\to\infty$ such that $\epsilon/b\to 0$; also assume that $n_b/n \xrightarrow{p} 0$. Note that the matrix $\mathbf{A}_{p\times k}$ under the constraint $\mathbf{A}^\top\mathbf{A}=\mathbf{I}$ is not unique, but the subspace spanned by the columns of $\mathbf{A}$, $\mathcal{S}(\mathbf{A})$, is unique and therefore, for simplicity assume that $\mathbf{A}$ is unique.

Using proof by contradiction, suppose that $\widehat{\mathbf{A}}$ does not converge to $\mathbf{A}$ almost surely (a.s.). Then there exists a subsequence, still denoted by n, and a matrix $\mathbf{A}_0$ satisfying the constraints $\mathbf{A}_0^\top \mathbf{A}_0 = \mathbf{I}$ such that $\widehat{\mathbf{A}} \to \mathbf{A}_0$ *a.s.*, where $\mathbf{A} \neq \mathbf{A}_0$. Therefore, for any $\epsilon > 0$ and n large enough, from Lemma 1:

$$\begin{aligned} f_n\left(y_i\right) &= f\left(y_i\right) + \delta_{1,i} \\ f_n\left(\widehat{\mathbf{A}}^\top \mathbf{x}_i\right) &= f\left(\widehat{\mathbf{A}}^\top \mathbf{x}_i\right) + \Delta_{1,i} = f\left(\mathbf{A}_0^\top \mathbf{x}_i\right) + \delta_{2,i} \\ f_n\left(y_i, \widehat{\mathbf{A}}^\top \mathbf{x}_i\right) &= f\left(y_i, \widehat{\mathbf{A}}^\top \mathbf{x}_i\right) + \Delta_{2,i} = f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right) + \delta_{3,i}, \end{aligned}$$

such that $|\Delta_{j,i}|, |\delta_{j,i}| < \epsilon$ for all $i = 1, 2, \ldots, n$ and $j \in \{1, 2, 3\}$. The first equalities in the last two equations follow from the conclusion of Lemma 1, and the remaining equalities from the uniform continuity conditions. Taking the natural log of each equation above gives:

$$\begin{aligned} \ln\left\{ f_n\left(y_i\right) \right\} &= \ln\left\{ f\left(y_i\right) \right\} + \ln\left\{ 1 + \delta_{1,i}/f\left(y_i\right) \right\} \\ \ln\left\{ f_n\left(\widehat{\mathbf{A}}^\top \mathbf{x}_i\right) \right\} &= \ln\left\{ f\left(\mathbf{A}_0^\top \mathbf{x}_i\right) \right\} + \ln\left\{ 1 + \delta_{1,i}/f\left(\mathbf{A}_0^\top \mathbf{x}_i\right) \right\} \\ \ln\left\{ f_n\left(y_i, \widehat{\mathbf{A}}^\top \mathbf{x}_i\right) \right\} &= \ln\left\{ f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right) \right\} + \ln\left\{ 1 + \delta_{1,i}/f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right) \right\}. \end{aligned}$$

Subtracting the first two equations from the last gives,

$$\ln\left\{ \frac{f_n(y_i, \widehat{\mathbf{A}}^\top \mathbf{x}_i)}{f_n\left(y_i\right) f_n(\widehat{\mathbf{A}}^\top \mathbf{x}_i)} \right\} = \ln\left\{ \frac{f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right)}{f\left(y_i\right) f\left(\mathbf{A}_0^\top \mathbf{x}_i\right)} \right\} + \ln\left\{ \frac{1 + \delta_{3,i}/f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right)}{\{1 + \delta_{1,i}/f\left(y_i\right)\}\{1 + \delta_{2,i}/f\left(\mathbf{A}_0^\top \mathbf{x}_i\right)\}} \right\}.$$

Next, letting $G_{1,i} = \frac{f_n(y_i, \widehat{\mathbf{A}}^\top \mathbf{x}_i)}{f_n(y_i) f_n(\widehat{\mathbf{A}}^\top \mathbf{x}_i)}$, $G_{2,i} = \frac{f(y_i, \mathbf{A}_0^\top \mathbf{x}_i)}{f(y_i) f(\mathbf{A}_0^\top \mathbf{x}_i)}$, $w_{1,i} = \left\{ 1 + \delta_{1,i}/f\left(y_i\right) \right\}$, $w_{2,i} = \left\{ 1 + \delta_{2,i}/f\left(\mathbf{A}_0^\top \mathbf{x}_i\right) \right\}$ and $w_{3,i} = \left\{ 1 + \delta_{3,i}/f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right) \right\}$, eliminating the natural logarithms, and raising to the $\alpha - 1$ power,

$$\begin{aligned} &(G_{1,i})^{\alpha-1} = (G_{2,i})^{\alpha-1} \left(\frac{w_{3,1}}{w_{1,i} w_{2,i}} \right)^{\alpha-1} \implies \\ &\qquad \ln\left\{ \frac{1}{n} \sum_{i=1}^{n} (G_{1,i})^{\alpha-1} \right\} = \ln\left\{ \frac{1}{n} \sum_{i=1}^{n} (G_{2,i})^{\alpha-1} \left(\frac{w_{3,1}}{w_{1,i} w_{2,i}} \right)^{\alpha-1} \right\}. \end{aligned} \tag{16}$$

Under the restriction that $i \in \chi_b$, and since $|\delta_{j,i}| < \epsilon$ and $\epsilon/b \to 0$ by definition, $\left\{ \delta_{1,i}/f\left(y_i\right) \right\}$, $\left\{ \delta_{2,i}/f\left(\mathbf{A}_0^\top \mathbf{x}_i\right) \right\}$ and $\left\{ \delta_{3,i}/f\left(y_i, \mathbf{A}_0^\top \mathbf{x}_i\right) \right\} \to 0$ as $n \to \infty$. This implies that $\lambda_i = (w_{3,1}/w_{1,i} w_{2,i})^{\alpha-1} \to 1$ as $n \to \infty$. Define, $\lambda_n^{max} = \max\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ and $\lambda_n^{min} = \min\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. These results together with (16) provide the inequality,

$$\begin{aligned} \widehat{\mathcal{R}}_\alpha^b(\widehat{\mathbf{A}}) &= \frac{1}{\alpha - 1} \ln\left[\frac{1}{n} \sum_{i=1}^{n} J\left(i \in \chi_b\right) \left\{ \frac{f(y_i, \mathbf{A}_0^\top \mathbf{x}_i)}{f(y_i) f(\mathbf{A}_0^\top \mathbf{x}_i)} \right\}^{\alpha-1} \lambda_i \right] \\ &\leq \frac{1}{\alpha - 1} \ln\left[\frac{1}{n} \sum_{i=1}^{n} J\left(i \in \chi_b\right) \left\{ \frac{f(y_i, \mathbf{A}_0^\top \mathbf{x}_i)}{f(y_i) f(\mathbf{A}_0^\top \mathbf{x}_i)} \right\}^{\alpha-1} \lambda_n^{max} \right] \\ &= \mathcal{R}_{n,\alpha}^b(\mathbf{A}_0) + \{1/(\alpha - 1)\} \ln(\lambda_n^{max}) \\ &= \mathcal{R}_{n,\alpha}^b(\mathbf{A}_0) + o(1). \end{aligned}$$

Substituting λ_n^{min} for λ_i above, and following an analogous argument, then $\widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) \geq \mathcal{R}^b_{n,\alpha}(\mathbf{A}_0)+\{1/(\alpha-1)\}\ln(\lambda_n^{min}) = \mathcal{R}^b_{n,\alpha}(\mathbf{A}_0)+o^*(1)$. Next, letting χ_b^c denote the complement of χ_b, we have that

$$\mathcal{R}^b_{n,\alpha}(\mathbf{A}_0) + o^*(1) \leq \widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) \leq \mathcal{R}^b_{n,\alpha}(\mathbf{A}_0) + o(1) \text{ and } \mathcal{R}^b_{n,\alpha}(\mathbf{A}_0) = \mathcal{R}_{n,\alpha}(\mathbf{A}_0) - \mathcal{R}^{b^c}_{n,\alpha}(\mathbf{A}_0).$$

Substituting the expression on the right into the left, and subtracting $\mathcal{R}(\mathbf{A}_0)$, results in the inequality,

$$\begin{aligned}\mathcal{R}_{n,\alpha}(\mathbf{A}_0) - \mathcal{R}_\alpha(\mathbf{A}_0) - \mathcal{R}^{b^c}_{n,\alpha}(\mathbf{A}_0) + o^*(1) \leq \widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) - \mathcal{R}_\alpha(\mathbf{A}_0)&\\ \leq \mathcal{R}_{n,\alpha}(\mathbf{A}_0) - \mathcal{R}_\alpha(\mathbf{A}_0) - \mathcal{R}^{b^c}_{n,\alpha}(\mathbf{A}_0) + o(1).&\end{aligned}$$

Next, by the law of large numbers $\mathcal{R}_{n,\alpha}(\mathbf{A}_0) - \mathcal{R}_\alpha(\mathbf{A}_0) \to 0$ as $n \to \infty$, and $\mathcal{R}^{b^c}_{n,\alpha}(\mathbf{A}_0) \to 0$ since $n_b/n \to 0$ as $n \to \infty$, and therefore,

$$\lim_{n\to\infty} \widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) = \mathcal{R}_\alpha(\mathbf{A}_0).$$

Now, by assumption: $\widehat{\mathbf{A}} = \operatorname{argmax} \widehat{\mathcal{R}}_\alpha(\mathbf{A}^*)$ [i], $\mathbf{A} = \operatorname{argmax} \mathcal{R}_\alpha(\mathbf{A}^*)$ [ii], and $\mathcal{R}_\alpha(\mathbf{A}_0) \leq \mathcal{R}_\alpha(\mathbf{A})$ [iii]. Therefore, [i] implies $\widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) \geq \widehat{\mathcal{R}}^b_\alpha(\mathbf{A}) \implies \mathcal{R}_\alpha(\mathbf{A}_0) = \lim_{n\to\infty} \widehat{\mathcal{R}}^b_\alpha(\widehat{\mathbf{A}}) \geq \lim_{n\to\infty} \widehat{\mathcal{R}}^b_\alpha(\mathbf{A}) = \mathcal{R}_\alpha(\mathbf{A})$ [iv]. Then, [iii] and [iv] imply that $\mathcal{R}_\alpha(\mathbf{A}) \leq \mathcal{R}_\alpha(\mathbf{A}_0) \leq \mathcal{R}_\alpha(\mathbf{A}) \implies \mathcal{R}_\alpha(\mathbf{A}) = \mathcal{R}_\alpha(\mathbf{A}_0)$, which contradicts the assumed uniqueness of $\mathbf{A}$ and therefore, $\widehat{\mathbf{A}} \to \mathbf{A}$ almost surely.

A.4 Proof of Proposition 3

To be more consistent with the influence function literature, and to ease in the exposition of the proof of Proposition 3, we use the differential notation $dF(y, \mathbf{A}^\top\mathbf{x}) = f(y, \mathbf{A}^\top\mathbf{x})d(\mathbf{A}^\top\mathbf{x})dy$ in (2). The regularity conditions allowing the interchange of differentiation and integration is assumed in the following proof.

Proof of Proposition 3:

As defined in (6), $T(F)$ and $T(F_\varepsilon)$ maximize (2) for the distribution functions F and F_ε, respectively, and therefore, each functional correspondingly solve

$$\int_y \int_{T(F)^\top\mathbf{x}} \dot{S}_\alpha(T(F); \mathbf{w})\ dF(y, T(F)^\top\mathbf{x}) = 0 \tag{17}$$

and

$$\int_y \int_{T(F_\varepsilon)^\top\mathbf{x}} \dot{S}_\alpha(T(F_\varepsilon); \mathbf{w})\ dF_\varepsilon(y, T(F_\varepsilon)^\top\mathbf{x}) = 0. \tag{18}$$

Rewriting the contamination distribution as $F_\varepsilon = F + \varepsilon(\Delta_{\mathbf{w}_0} - F)$ and substituting into (18) yields the expression,

$$\begin{aligned}&\int_y \int_{T(F_\varepsilon)^\top\mathbf{x}} \dot{S}_\alpha(T(F_\varepsilon); \mathbf{w})\ dF(y, T(F_\varepsilon)^\top\mathbf{x})\\ &\qquad + \varepsilon \int_y \int_{T(F_\varepsilon)^\top\mathbf{x}} \dot{S}_\alpha(T(F_\varepsilon); \mathbf{w})\ d(\Delta_{\mathbf{w}_0} - F)(y, T(F_\varepsilon)^\top\mathbf{x}) = 0.\end{aligned} \tag{19}$$

Next, differentiating the expression in (19) with respect to ε,

$$\begin{aligned}\frac{\partial}{\partial\varepsilon}\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}} &\dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ dF(y,T(F_\varepsilon)^\top\mathbf{x})\\ &+\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}} \dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ d(\Delta_{\mathbf{w}_0}-F)(y,T(F_\varepsilon)^\top\mathbf{x})\\ &\qquad+\ \varepsilon\ \frac{\partial}{\partial\varepsilon}\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}} \dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ d(\Delta_{\mathbf{w}_0}-F)(y,T(F_\varepsilon)^\top\mathbf{x})=0. \qquad (20)\end{aligned}$$

Taking the partial derivative of (20) with respect to ε and evaluating at $\varepsilon=0$ results in the equality,

$$\begin{aligned}\left\{\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}}\frac{\partial}{\partial\varepsilon}\dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ dF(y,T(F_\varepsilon)^\top\mathbf{x})\right\}\Bigg|_{\varepsilon=0}&\times\frac{\partial}{\partial\varepsilon}T(F_\varepsilon)\ \Bigg|_{\varepsilon=0}\\ +\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}}&\dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ d\Delta_{\mathbf{w}_0}\\ &-\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}}\dot{S}_\alpha(T(F);\mathbf{w})\ dF(y,T(F)^\top\mathbf{x})=0. \qquad (21)\end{aligned}$$

Noting that the last integral expression in (21) equals 0 since $T(F)$ solves (17), and then solving the equality for (7) yields the result,

$$\frac{\partial}{\partial\varepsilon}T(F_\varepsilon)\ \Bigg|_{\varepsilon=0}=-\left[\left\{\int_y\int_{T(F_\varepsilon)^\top\mathbf{x}}\frac{\partial}{\partial\varepsilon}\dot{S}_\alpha(T(F_\varepsilon);\mathbf{w})\ dF(y,T(F_\varepsilon)^\top\mathbf{x})\right\}\Bigg|_{\varepsilon=0}\ \right]^{-1}\times\dot{S}(T(F_\varepsilon);\mathbf{w}_0).$$

A.5 Rényi properties

The limiting result, $\lim_{\alpha\to1}\mathcal{R}_\alpha(\mathbf{A})=D_{KL}(\mathbf{A})=D_{KL}\{f(Y,\mathbf{A}^\top\mathbf{X})||f(Y)f(\mathbf{A}^\top\mathbf{X})\}$, was established in Erven and Harremoës [13], and can be shown by directly taking the limit, applying L'Hospital's rule, of the integral expression of the index in (2).

The bound $\mathcal{R}_\alpha(\mathbf{A})\le D_{KL}(\mathbf{A})$ is shown by first considering $h(u)=\ln(u), u>0$, then $h''(u)=-1/u^2<0 \implies h(u)$ is concave for all u and thus, for a positive random variable $\mathbf{U}$, $\mathrm{E}\{h(\mathbf{U})\}\le h\{\mathrm{E}(\mathbf{U})\}$ by Jensen's inequality. Then, with $\mathbf{U}=f(Y,\mathbf{A}^\top\mathbf{X})/f(Y)f(\mathbf{A}^\top\mathbf{X})$,

$$\begin{aligned}\mathrm{E}\left[\ln\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha-1}\right]&\le\ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha-1}\right]\\ \implies\qquad(\alpha-1)\mathrm{E}\left[\ln\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}\right]&\le\ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha-1}\right]\\ \implies\qquad\mathrm{E}\left[\ln\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}\right]&\ge\frac{1}{\alpha-1}\ \ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha-1}\right].\end{aligned}$$

The monotonicity, $\mathcal{R}_{\alpha_1}(\mathbf{A})\le\mathcal{R}_{\alpha_2}(\mathbf{A})$ when $0<\alpha_1<\alpha_2<1$, can be established in an analogous argument, by noting that for $h(u)=u^{\left(\frac{\alpha_1-1}{\alpha_2-1}\right)}, u>0, h''(u)=\left(\frac{1-\alpha_1}{1-\alpha_2}\right)\left(\frac{1-\alpha_1}{1-\alpha_2}-1\right)u^{\left(\frac{\alpha_1-1}{\alpha_2-1}\right)-2}>0$, since $\frac{1-\alpha_1}{1-\alpha_2}>1 \implies h(u)$ is convex for all u. Then, for the previously defined $\mathbf{U}$, by

Jensen's inequality

$$\begin{aligned}
&\mathrm{E}\left[\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_2-1}\right]^{\frac{\alpha_1-1}{\alpha_2-1}} \geq \left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_2-1}\right]^{\frac{\alpha_1-1}{\alpha_2-1}}\\
&\qquad \implies \ln\left(\mathrm{E}\left[\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_2-1}\right]^{\frac{\alpha_1-1}{\alpha_2-1}}\right) \geq \left(\frac{\alpha_1-1}{\alpha_2-1}\right)\ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_2-1}\right]\\
&\qquad \implies \frac{1}{\alpha_1-1}\ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_1-1}\right] \leq \frac{1}{\alpha_2-1}\ln\left[\mathrm{E}\left\{\frac{f(Y,\mathbf{A}^\top\mathbf{X})}{f(Y)f(\mathbf{A}^\top\mathbf{X})}\right\}^{\alpha_2-1}\right].
\end{aligned}$$

A.6 Hellinger-Bhattacharyya distance equivalence

Consider the Hellinger-Bhattacharyya distance

$$HB = \left[\int_y\int_{\mathbf{A}^\top\mathbf{x}}\left\{\sqrt{f(y,\mathbf{A}^\top\mathbf{x})} - \sqrt{f(y)f(\mathbf{A}^\top\mathbf{x})}\ \right\}^2 d(\mathbf{A}^\top\mathbf{x})\ dy\right]^{1/2}.$$

Then,

$$\begin{aligned}
(HB)^2 &= \int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[f(y,\mathbf{A}^\top\mathbf{x})^{1/2} - \{f(y)f(\mathbf{A}^\top\mathbf{x})\}^{1/2}\right]^2 d(\mathbf{A}^\top\mathbf{x})\ dy\\
&= \int_y\int_{\mathbf{A}^\top\mathbf{x}}\left[f(y,\mathbf{A}^\top\mathbf{x}) + f(y)f(\mathbf{A}^\top\mathbf{x}) - 2f(y,\mathbf{A}^\top\mathbf{x})^{1/2}\{f(y)f(\mathbf{A}^\top\mathbf{x})\}^{1/2}\right] d(\mathbf{A}^\top\mathbf{x})\ dy\\
&= \int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y,\mathbf{A}^\top\mathbf{x})\ d(\mathbf{A}^\top\mathbf{x})\ dy + \int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y)f(\mathbf{A}^\top\mathbf{x})\ d(\mathbf{A}^\top\mathbf{x})\ dy\\
&\qquad\qquad - 2\int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y,\mathbf{A}^\top\mathbf{x})^{1/2}\{f(y)f(\mathbf{A}^\top\mathbf{x})\}^{1/2}\ d(\mathbf{A}^\top\mathbf{x})\ dy\\
&= 2\left[1 - \int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y,\mathbf{A}^\top\mathbf{x})^{1/2}\{f(y)f(\mathbf{A}^\top\mathbf{x})\}^{1/2}\ d(\mathbf{A}^\top\mathbf{x})\ dy\right],
\end{aligned}$$

since $\int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y,\mathbf{A}^\top\mathbf{x})\ d(\mathbf{A}^\top\mathbf{x})\ dy = \int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y)f(\mathbf{A}^\top\mathbf{x})\ d(\mathbf{A}^\top\mathbf{x})\ dy = 1$. Therefore, $-2\ln\{1-(HB)^2/2\} = -2\ln\{\int_y\int_{\mathbf{A}^\top\mathbf{x}} f(y,\mathbf{A}^\top\mathbf{x})^{1/2}\ \{f(y)f(\mathbf{A}^\top\mathbf{x})\}^{1/2}\ d(\mathbf{A}^\top\mathbf{x})\ dy\} = D_{1/2}\{f(Y,\mathbf{A}^\top\mathbf{X})||f(Y)\ f(\mathbf{A}^\top\mathbf{X})\} = \mathcal{R}_{1/2}(\mathbf{A})$.

References

[1] A. Basu, I. Harris, N. Hjort, M. Jones, Robust and efficient estimation by minimising a density power divergence, Biometrika 85 (1998) 549—599.

[2] J. Bénasséni, Sensitivity coefficients for the subspaces spanned by principal components, Comm. Statist.: Theory and Methods, 19 (1990) 2021–2034.

[3] P. Čížek, W. Härdle, Robust Adaptive Estimation of Dimension Reduction Space SBF 373, Berlin: Humboldt University of Berlin. (Discussion Paper) 1 (2003).

[4] P. Čížek , Robust estimation of dimension reduction space, Proceedings in Computational Science. Antoch, J. (ed.). Heidelberg: Physica-Verlag (2004) 871–878.

[5] P. Čížek, W. Härdle, Robust estimation of dimension reduction space, Comp. Statist. Data Anal. 51 (2006) 545-555.

[6] R.D. Cook, S. Weisberg, Discussion of Li, J. Amer. Statist. Assoc. 86 (1991) 328–332.

[7] R.D. Cook, X. Yin, Dimension-reduction and visualization in discriminant analysis (Invited with discussion), Australia & New Zealand Journal of Statistics 43 (2001) 147–200.

[8] R.D. Cook, Fisher lecture: Dimension reduction in regression (with discussion), Statist. Science 22 (2007) 1–26.

[9] N. Cressie, C. Read, Multinomial goodness-of-fit tests, Journal of the Royal Statistical Society: Series B 46 (1984) 440—464.

[10] F. Critchley, Influence in principal components analysis, Biometrika, 72 (1985) 627–636.

[11] F. Critchley, On preferred point geometry in statistics, J. Statist. Plan. Infer. 102 (2002) 229–245.

[12] Y. Dong, Z. Yu, L. Zhu, Robust inverse regression for dimension reduction, J. Multivariate Anal. 134 (2015) 71–81.

[13] T. Ervan, P. Harremoës, Rényi Divergence and Kullback-Leibler Divergence, IEEE Transactions on Information Theory 60 (2014) 3797–3820.

[14] U. Gather, T. Hilker, C. Becker, A robustified version of sliced inverse regression. In Statistics in Genetics and in the Environmental Sciences (L. T. Fernholz,S. Morgenthaler and W. Stahel, eds.) (2001) 147–157. Birkh¨auser, Basel.

[15] U. Gather, T. Hilker, C. Becker, A note on outlier sensitivity of sliced inverse regression, Statist. 36 (2002) 271–281.

[16] F.R. Hampel, E.M. Ronchetti, P.J. Rousseeuw, W.A. and Stahel, Robust Statistics. The Approach Based on Influence Functions. Wiley & Sons, New York 1986.

[17] W. Härdle, A. Tsybakov, "How sensitive are average derivatives," CORE Discussion Papers 1991044, Universitë catholique de Louvain (1991).

[18] X. He, W.K. Fung, Discussion of "Dimension Reduction and Visualization in Discriminant Analysis" by Cook and Yin, Australian & New Zealand Journal of Statistics, 43 (2001) 190–193.

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

[20] R. Iaci, T.N. Sriram, X. Yin, Multivariate Association and Dimension Reduction: A Generalization of Canonical Correlation Analysis, Biometrics 66 (2010) 1107–1118.

[21] R. Iaci, T.N. Sriram, Robust multivariate association and dimension reduction using density divergences, J. Multivariate Anal. 117 (2013) 281–295.

[22] R. Iaci, X Yin, L. Zhu, The Dual Central Subspaces in dimension reduction, J. Multivariate Anal. 145 (2016) 178–189.

[23] J. Kiefer, On large deviations of the empiric d.f. of vector chance variables and a law of the iterated logarithm, Pacific J. Math 11 (1961) 649–659.

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

[25] L. A. Prendergast, J. A. Smith, Influence Functions for Dimension Reduction Methods: An Example Influence Study of Principal Hessian Direction Analysis, Scandinavian J. Statist.: theory and applications, 37 (20100 588–611.

[26] A. Rényi, On measures of information and entropy, Proceedings of the fourth Berkeley Symposium on Mathematics, Statistics and Probability (1961) 547–561.

[27] P.J. Rousseeuw, A.M. Leroy. Robust Regression and Outlier Detection, John Wiley & Sons, New York, (2003).

[28] L. Ruschendorf, Consistency of estimators for multivariate density functions and for the mode. Sankhya. A, 39 (1977) 243–50.

[29] D.W. Scott, Multivariate density estimation: Theory, Practice and Visualization, John Wiley & Sons, New York, (1992).

[30] R.G. Staudte, S.J. Sheather, Robust Estimation and Testing. Wiley, New York 1990.

[31] B.W. Silverman, Density estimation for statistics and data analysis, Chapman & Hall, 1986.

[32] Q. Wang, X. Yin, F. Critchley, Dimension reduction based on Hellinger integral, Biometrika, 102 (2015) 95–106.

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

[34] Y. Xue, N. Zhang, X. Yin, H. Zeng, Sufficient dimension reduction using Hilbert-Schmidt independence criterion, Comp. Statist. & Data Analysis 115 (2017) 67–78.

[35] Y. Xue, Q. Wang, X. Yin, A unified approach to sufficient dimension reduction, J. Statist. Planning and Inference 197 (2018) 168–179.

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

[37] V. J. Yohai, M. E. Szretter Noste, A Robust Proposal for Sliced Inverse Regression, (2005).

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

[39] X. Yin, T.N. Sriram, Common canonical variates for independent groups using information theory, Statist. Sinica 18 (2008) 335–353.

[40] J. Zhang, X. Chen, Robust sufficient dimension reduction via ball covariance, Comp. Statist. & Data Analysis 140 (2019) 144–154.

[41] X. Zhang, Q. Mai, H. Zou. The Maximum Separation Subspace in Sufficient Dimension Reduction with Categorical Response, J. Machine Learning Research 21 (2020) 1–40.

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