Sankhyā A : The Indian Journal of Statistics
DOI 10.1007/s13171-017-0102-x
© 2017, Indian Statistical Institute

The Effect of Data Contamination in Sliced Inverse Regression and Finite Sample Breakdown Point

Ulrike Genschel
Iowa State University, Ames, USA

Abstract

Dimension reduction procedures have received increasing consideration over the past decades. Despite this attention, the effect of data contamination or outlying data points in dimension reduction is, however, not well understood, and is compounded by the issue that outliers can be difficult to classify in the presence of many variables. This paper formally investigates the influence of data contamination for sliced inverse regression (SIR), which is a prototypical dimension reduction procedure that targets a lower-dimensional subspace of a set of regressors needed to explain a response variable. We establish a general theory for how estimated reduction subspaces can be distorted through both the number and direction of outlying data points. The results depend critically on the regressor covariance structure and the most harmful types of data contamination are shown to differ in cases where this covariance structure is known or unknown. For example, if the covariance structure is estimated, data contamination is proven to produce an estimated subspace that is automatically orthogonal to the directions of outlying data points, constituting a potentially serious loss of information. Our main results demonstrate the degree to which data contamination indeed causes incorrect dimension reduction, depending on the amount, magnitude, and direction of contamination. Further, by metricizing distances between dimension reduction subspaces, worst case results for data contamination can be formulated to define a finite sample breakdown point for SIR as a measure of global robustness. Our theoretical findings are illustrated through simulation.

AMS (2000) *subject classification.* Primary 62G35, Secondary 62G08.
Keywords and phrases. Subspace estimation, Data contamination, Sliced inverse regression, Spectral decomposition, Breakdown

1 Introduction

Dimension reduction is an important statistical problem, which continues to receive natural consideration in an era where increasingly complex and

Electronic supplementary material The online version of this article (doi:10.1007/s13171-017-0102-x) contains supplementary material, which is available to authorized users.

large data structures are observed, particularly in regression problems (cf. Cook and Ni, 2005, 2006; Li, 2007; Li and Dong, 2009; Chen et al., 2010; Li et al., 2010; Ma and Zhu, 2012, 2013 or Yin and Hilafu, 2014 among many others). Given a set of p regressors, $\mathbb{X} \in \mathbb{R}^p$, and a response $Y \in \mathbb{R}$, dimension reduction may target a lower dimensional regressor subspace, say $\mathcal{B} \subset \mathbb{R}^p$, for appropriately explaining the response. $\mathcal{B}$ then becomes the basis for estimating the functional relationship between $\mathcal{B}$ and the response Y. Despite the widespread use of dimension reduction procedures, little attention has been paid to developing a better theoretical understanding of the effect of data contamination on such subspace estimates. This aspect motivates a theoretical and quantitative investigation of the robustness properties of one dimension reduction procedure in the following.

Sliced inverse regression, or SIR for short, was first introduced by Li (1991), and is one of several classical and well-known procedures that aim at estimating a so-called *sufficient dimension reduction subspace*. Subsequent procedures were proposed, for example, by Li (1991, 1992), Cook (2000), Cook and Weisberg (1991), Cook and Yin (2000), Cook and Ni (2005, 2006), Xia et al. (2002), and more recently by Li et al. (2010), Ma and Zhu (2012), Forzani et al. (2012) and Yin and Hilafu (2014). Reliable inference about the response often depends on identification of a subspace $\mathcal{B}$, including its dimension $\mathcal{K}$. Because of this, sensitivity of dimension reduction procedures, like SIR, to data contamination is important to understand. Although research, with particular consideration of SIR, has been done in this area (cf. Li, 1991; Hilker, 1997; Bond, 1999; Cook and Critchley, 2000; Becker, 2001; Gather et al., 2002; Prendergast, 2005, 2006, 2007), the results are incomplete, applying mostly to the single-index model only and, at times, even seemingly inconsistent. In particular, SIR is a complex multi-step procedure, involving evaluation of two location and two scatter functionals. While its estimation steps have been investigated separately in terms of robustness (Hilker, 1997; Becker, 2001; Gather et al., 2002), the global robustness of the procedure as a whole has not been theoretically examined. Our goal is hence to provide a theoretical and quantitative study of the effects of data contamination on SIR in subspace $\mathcal{B}$ estimation (of possibly arbitrary dimension $\mathcal{K}$). Although dimension reduction methodology can differ depending on the underlying data structure and motivation, the insights derived from the study of the sensitivity of SIR to outliers could help towards a better understanding of the robustness of other dimension reduction procedures as well.

To place our work against existing literature, note that the sensitivity of SIR to outliers has been investigated in varying manners by several authors. As outliers can easily remain undetected in high dimensional data settings (Rousseeuw and Leroy, 1987) concerns about the effects of data contamination are valid and this issue requires further attention. Li (1991), for example, suggested that SIR can be robust against outlying observations and Bond (1999), in investigating the robustness of SIR to violations of distributional assumptions, concluded that SIR is “fairly robust, not being affected by even fairly sizable perturbations when the response function was well behaved.” Cook and Critchley (2000) stated that outlying observations only result in subspace estimates having additional linear combinations to the true subspace. To the contrary, Sheather and McKean (1997), Welsh (2001), Gather et al. (2002), and Predergast (2005, 2006, 2007) have mentioned potential sensitivity of SIR to outliers. In fact, early results combined from Sheather and McKean (1997) and Sheather and McKean (2001) hint that outliers can affect SIR quite differently. Specifically, while Sheather and McKean (1997) indicated a sensitivity of SIR to outliers, Sheather and McKean (2001) give a data example (Diabetes in Pima Indians, Ripley, 1996) where removing two outliers has only a small effect on the estimated e.d.r. direction. Hilker (1997), Becker (2001) and Gather et al. (2002), focusing on the case $\mathcal{K} = 1$, have argued that the direction of extreme data points can indeed affect SIR. This finding is supported by Prendergast (2005, 2006), who extensively studied the influence function of SIR, also noting that the influence of outliers on SIR differs and can even be zero depending on the direction. In response to the sensitivity of SIR to outliers, Sheather and McKean (1997), Gather et al. (2001), Dong et al. (2015) and Chiancone et al. (2016) have proposed robustified alternatives of SIR. In addition to outliers, Dong et al. (2015) also address the effect of inliers on SIR, which have received even less attention in the literature. We believe that such different findings are due to the fact that the type of data contamination plays a decisive role in the robustness of SIR. That is, single outlying data points may not necessarily yield a bad estimate of a dimension reduction subspace $\mathcal{B} \subset \mathbb{R}^{\mathcal{K}}$ simply due to their nature of being extreme. Our results prove, however, that the *amount* and *directions* of extreme data points can dramatically influence the estimation of $\mathcal{B}$. In fact, we show that data contamination is possible such that it results in a total loss of all important linear combinations of the subspace as well as the inclusion of false (e.g., completely orthogonal) information. Our results expand upon the findings of Hilker (1997), Becker (2001), and Gather et al. (2002) that the direction of an outlying data point is important, but we further quantify the effect of the amount of data contamination and our findings also apply in the estimation of a general dimension reduction subspace $\mathcal{B} \subset \mathbb{R}^{\mathcal{K}}$ (which need not have dimension $\mathcal{K} = 1$ as considered by Hilker (1997), Becker (2001), and Gather et al. (2002) above).

Note that SIR involves a type of principal component analysis to identify or estimate vectors (i.e., directions) to provide the most important linear combinations of the regressor variables $\mathbb{X} \in \mathbb{R}^p$ for explaining the response Y (see Section 2.1). This process involves the regressor covariance $\mathrm{Var}(\mathbb{X}) = \Sigma \in \mathbb{R}^{p \times p}$ structure (and mean $\mathrm{E}\mathbb{X} = \mu \in \mathbb{R}^p$), which is either known or unknown requiring estimation. Robustness properties of Σ, such as the breakdown point, qualitative robustness or the influence function (c.f., Li and Chen, 1985; Croux and Haesbroeck, 2000; or Croux and Ruiz-Gazen, 2005, for example) have been established in the literature and are well known. Tyler (2005) illustrates through an example that "the breakdown point of the robust covariance matrix has no information regarding the principal component vectors, since the breakdown of a covariance matrix only implies that either the largest root can become arbitrarily large or the smallest root can become arbitrarily small." Thus, the existing knowledge on the sensitivity of a sample covariance matrix does not naturally translate to the same results on the sensitivity of SIR. For more details we refer to the discussion by Tyler (2005).

To highlight the findings for data contamination, we may characterize the main results as follows. When Σ is known, the effects of contamination can be precisely quantified upon replacing k data points with contaminated values where the choice of k controls the level of contamination. In particular, if these contaminated points are placed in k distinct directions of contamination, say $\{\widetilde{\beta}_h\}_{h=1}^k \subset \mathbb{R}^p$, which are *orthogonal* to the true dimension reduction subspace $\mathcal{B} \subset \mathbb{R}^{\mathcal{K}}$, then a contaminated subspace estimate $\widehat{\mathcal{B}}$ will lose information and become orthogonal to any part of $\mathcal{B}$ that is orthogonal to $\{\widetilde{\beta}_h\}_{h=1}^k$. On the other hand, the effects of data contamination differ dramatically when the regressor covariance Σ is unknown. In this case, *whatever* directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^k$ are used, the resulting contaminated subspace estimate $\widehat{\mathcal{B}}$ will be orthogonal to these directions. Thus, contrary to the Σ known situation, if data contamination lies in the *same direction* as vectors spanning the true reduction subspace $\mathcal{B}$, the contaminated estimate $\widehat{\mathcal{B}}$ will fail to capture important information about $\mathcal{B}$. While the known regressor moments (Σ, μ) are admittedly less common, we include results for the Σ known case due to the dramatic differences in contamination necessary to cause a worst possible estimate $\widehat{\mathcal{B}}$ as well as due to their potential insight about effect of outliers on $\widehat{\mathcal{B}}$ if we were able to estimate μ and Σ sufficiently well, e.g., by using adequately robust estimators. Whether Σ is known or not, one can metricize the distance between subspaces of $\mathbb{R}^p$ and examine worst case estimation under data contamination, which, in turn, allows us to propose a definition of the finite sample breakdown point (amount of data contamination for producing an estimated subspace that is maximally distant from the true or target reduction subspace $\mathcal{B}$). Several results can be formulated about such worst case estimation, which intricately depend on a combination of factors such as the dimension p of the regressors, the dimension $\mathcal{K}$ of the target subspace $\mathcal{B}$, and whether $\mathcal{K}$ requires estimation or not.

The remainder of the manuscript is organized as follows. In Section 2, we describe the SIR dimension reduction procedure and outline the general data contamination scheme. Section 3 provides the main results on data contamination in dimension reduction, providing an overall characterization of the effects of k-point data contamination on subspace estimation. Section 4 applies these findings in a further study of worst case estimation from contamination. Simulations in Section 5 numerically and concretely illustrate our theoretical findings on data contamination, and Section 6 offers concluding remarks. The proofs of the main results appear in a Supplement (Genschel, 2017).

2 Data Contamination for Corrupting Dimension Reduction

We first provide a brief description of the dimension reduction procedure, SIR (sliced inverse regression), and refer to Li (1991) amongst others for further reading. In Section 2.2, we describe a general data contamination pattern that, if present in the data, can corrupt SIR in various manners.

2.1. Background of SIR. Let $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, denote a random sample from $(\mathbb{X}, Y) \in \mathbb{R}^p \times \mathbb{R}$ with Y as the response variable of interest and $\mathbb{X} = (X_1, \ldots, X_p)^\top$ denoting a set of regressors. Further, assume that $Y = g(\mathbb{X}, \varepsilon)$, ε independent of $\mathbb{X}$. SIR aims to reduce the dimension of the regressor space by identifying $\mathcal{K} < p$ linear combinations $B^\top\mathbb{X}$ of $\mathbb{X}$, where B denotes a $p \times \mathcal{K}$ matrix with columns $\beta_1, \ldots \beta_{\mathcal{K}} \in \mathbb{R}^p$. Under model regularity conditions (c.f. Li, 1991), these essential linear combinations, $B^\top\mathbb{X}$, correspond to those needed to linearly span and completely explain the inverse regression curve $\mathrm{E}(\mathbb{X}|Y)$ in $\mathbb{R}^p$. The columns of B are identified through principal component analysis of the covariance matrix $V = \mathrm{Cov}[\mathrm{E}(\mathbb{Z}|Y)]$ of the conditional expectation $\mathrm{E}(\mathbb{Z}|Y)$ where $\mathbb{Z} = \Sigma^{-1/2}(\mathbb{X} - \mathrm{E}(\mathbb{X}))$ denotes the standardized regressor vector with $\Sigma = \mathrm{Cov}(\mathbb{X})$. The spectral decomposition of V identifies the $\mathcal{K}$ directions $\eta_1, \ldots, \eta_{\mathcal{K}}$ with highest variability in $\mathrm{E}(\mathbb{Z}|Y)$

as the eigenvectors of V corresponding to the largest eigenvalues of V. Undoing the standardization of η_i yields the actual directions $\beta_i = \Sigma^{-1/2}\eta_i$, $i = 1, \dots, \mathcal{K}$. While directions are unique only up to an orthogonal transformation, the corresponding $\mathbb{R}^p$-subspace $\mathcal{B} = \text{span}(\beta_1, \dots, \beta_\mathcal{K})$, however, is unique and is called the effective dimension reduction subspace; the vectors $\beta_1, \dots, \beta_\mathcal{K}$ are referred to as effective dimension reduction vectors or directions (henceforth denoted as the e.d.r. subspace and e.d.r. directions, respectively). Hence, of primary interest with SIR is the e.d.r. subspace $\mathcal{B}$.

2.2. Data Contamination. A crucial component of our results is that knowledge of the covariance and mean structure of the regressors, $\Sigma = \text{Cov}(\mathbb{X})$ and $\mu = \text{E}(\mathbb{X})$, (or lack of this knowledge) influences how contaminated observations are to be formulated in order to bend an estimate of the e.d.r. subspace $\mathcal{B}$ to a desired degree. As a result, two slightly different data contamination scenarios emerge, which we discuss next.

Suppose SIR is applied to a size n sample $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ of $(\mathbb{X}, Y)$ with $y_i \neq y_j \ \forall\ i \neq j$. Upon ordering $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ according to its y-values, i.e. $y_{(i)} < y_{(i+1)}$, the observations are divided into H slices $\mathbb{I}_h$, $h = 1, \dots, H$, defined as follows $y_{(1)}, \dots, y_{(n_1)} \in \mathbb{I}_1$; $y_{(n_1+1)}, \dots, y_{(n_1+n_2)} \in \mathbb{I}_2$; $\dots$; $y_{(n_1+\ldots+n_{H-1}+1)}, \dots, y_{(n)} \in \mathbb{I}_H$, where n_h denotes the sample size of the hth slice $\mathbb{I}_h$ and $\sum_{h=1}^H n_h = n$. For μ and Σ *known* (kn), an estimate of V (cf. Section 2.1) is given by

$$\widehat{V}_{\text{kn}} = \Sigma^{-1/2} \sum_{h=1}^{H} \widehat{p}_h (\overline{\mathbf{x}}_h - \mu)(\overline{\mathbf{x}}_h - \mu)^\top \Sigma^{-1/2} \qquad (1)$$

with $\overline{\mathbf{x}}_h$ and $\widehat{p}_h = n_h/n$ denoting the sample mean and proportion of the hth slice, $h = 1, \dots, H$. Alternatively, when Σ and μ are *unknown* (un.kn) and require estimation, respective estimates $\overline{\mathbf{x}} = \sum_{i=1}^n \mathbf{x}_i/n$ and $\widehat{\Sigma} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^\top/n$ are substituted for μ and Σ in (1) yielding $\widehat{V}_{\text{un.kn}}$. Spectral decompositions of $\widehat{V}_{\text{kn}}$ in (1) and $\widetilde{V}_{\text{un.kn}}$ return estimated e.d.r. directions $\widehat{\beta}_i = \Sigma^{-1/2}\widehat{\eta}_i$ and $\widehat{\beta}_i = \widehat{\Sigma}^{-1/2}\widehat{\eta}_i$, respectively, derived from the orthonormal eigenvectors $\widehat{\eta}_i$, $i = 1, \dots, \mathcal{K}$ corresponding to the $\mathcal{K}$ largest eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_\mathcal{K}$ of $\widehat{V}_{\text{kn}}$ or $\widetilde{V}_{\text{un.kn}}$; for notational simplicity, we use the same notation to denote eigenvalues $\widehat{\lambda}_i$ and eigenvectors $\widehat{\eta}_i$ in the decompositions of $\widehat{V}_{\text{kn}}$ and $\widetilde{V}_{\text{un.kn}}$, though these are not necessarily equal in both cases. For later results, we further note that, when the regressor covariance structure (e.g., Σ) is unknown, the estimated e.d.r. directions $\widehat{\beta}_1, \dots, \widehat{\beta}_\mathcal{K}$ can be determined directly as the eigenvectors corresponding to the $\mathcal{K}$ largest eigenvectors of

$$\widetilde{V}_{\text{un.kn}} = \widehat{\Sigma}^{-1} \sum_{h=1}^{H} \widehat{p}_h (\overline{\mathbf{x}}_h - \overline{\mathbf{x}})(\overline{\mathbf{x}}_h - \overline{\mathbf{x}})^\top; \qquad (2)$$

this is because $\widehat{V}_{\text{un.kn}}\widehat{\eta}_i = \widehat{\eta}_i\widehat{\lambda}_i$ if and only if $\widetilde{V}_{\text{un.kn}}\widehat{\beta}_i = \widehat{\beta}_i\widehat{\lambda}_i$. Finally, if the dimension $\mathcal{K}$ of the e.d.r. subspace is unknown, a step can be added to estimate this; we describe this component later in Section 4.

We exploit data slices in the contamination scheme to follow. Let $\{t_m\}_{m=1}^{\infty}$ denote a sequence of positive scaling factors such that $t_m \to \infty$ as $m \to \infty$. For each m and factor t_m, we create a contaminated sample, $(\mathbb{X}, Y)_m^{n,k}$, by replacing $k \leq H$ observations in the uncontaminated data $(\mathbb{X}, Y)^n$, where one observation is replaced by an outlying point in each of k distinct slices. Without loss of generality, we proceed as follows. For $h = 1, \ldots, H$ denote the index of the first observation in each slice $\mathbb{I}_h$ as $1_h = 1 + \sum_{i=1}^{h-1} n_i$ and suppose the first observation of each of the first $k \leq H$ slices $\mathbb{I}_1, \ldots, \mathbb{I}_k$ is replaced by arbitrary observations $(\widetilde{\mathbf{x}}_{m,1_h}, \widetilde{y}_{1_h}), h = 1, \ldots, k$, where response values $\{\widetilde{y}_{1_h} = y_{1_h}\}_{h=1}^{k}$ remain unchanged and thus observations assigned to slices $\mathbb{I}_h, h = 1, \ldots, H$ stay the same. The contamination of the $\mathbf{x}$-values is assumed to be of the following general structure

$$\widetilde{\mathbf{x}}_{m,1_h} = t_m\widetilde{\beta}_h + \widetilde{v}_h \in \mathbb{R}^p, \quad h = 1, \ldots, k, \tag{3}$$

for some vectors $\widetilde{\beta}_h, \widetilde{v}_h \in \mathbb{R}^p$, where we impose additional conditions on the directions $\{\widetilde{\beta}_h\}_{h=1}^{k}$ of contamination depending on whether or not the covariance structure of the regressors is known. For Σ and μ *known*, we select $\{\widetilde{\beta}_h\}_{h=1}^{k}$ to satisfy

$$\widetilde{\beta}_h\Sigma^{-1}\widetilde{\beta}_j^{\top} = I(h = j), \tag{4}$$

where $I(\cdot)$ denotes the indicator function (i.e., contaminated directions are orthonormal after a linear transformation). In the case of Σ and μ *unknown*, we require

$$\{\widetilde{\beta}_h\}_{h=1}^{k} \text{ are linearly independent.} \tag{5}$$

This provides a contaminated sample $(\mathbb{X}, Y)_m^{n,k}$ based on replacing the first observation in the first k slices. Note that in practice it might be more common to see multiple outlying observations in a single slice as opposed to being distributed across distinct slices. If all contamination essentially lies in the same direction, then our conclusions about the effect of outliers continue to hold. It becomes more difficult to quantify what happens if the directions of contamination substantially differ within the same slice. We suspect, then, that the relative magnitudes of outliers become a greater consideration, which requires and complicates further investigation.

The exact number k of contaminated points to use may be chosen in a way so that the resulting contaminated estimate of the e.d.r. subspace is maximally distant from the true one, $\mathcal{B}$; this topic is considered in Section 4. However, before this step, results can be formulated to show the general outcome of this data contamination scheme for SIR and that, in particular, one may concretely describe how such data contamination can corrupt an estimate of the e.d.r. subspace $\mathcal{B}$.

3 Main Results on Data Contamination and Subspace Estimation

In this section, we detail the effect various types of contamination have on SIR's ability to correctly estimate the e.d.r. subspace $\mathcal{B}$. We demonstrate that the degree to which an e.d.r. subspace estimate deviates from the true one, $\mathcal{B}$, is determined by the number of contaminated observations in distinct slices and the directions in which these observations are placed. That is, and contrary to many familiar statistical procedures, the magnitude of contamination (i.e., "how far out" a data point is relative to the rest of the data) does not determine the distortion of the subspace estimate, as much as the directions of contamination. Additionally, the overall effects of data contamination on e.d.r. subspace estimation, and the most serious directions of contamination, differ dramatically depending upon knowledge of the covariance matrix $\Sigma = \mathrm{Var}(\mathbb{X})$. Because of this, we separately treat cases of either known or unknown regressor covariance structures in Sections 3.1–3.2, respectively.

3.1. Results for Known Regressor Covariance. If $(\mathbb{X}, Y)_m^{n,k}$ denotes a contaminated sample from the k-slice contamination scheme from Section 2.2, $1 \leq k \leq \min\{p, H\}$, with known $\mathrm{E}(\mathbb{X}) = \mu \in \mathbb{R}^p$ and $\mathrm{Cov}(\mathbb{X}) = \Sigma \in \mathbb{R}^{p \times p}$, then let $\widehat{V}_{m,\mathrm{kn}}$ denote the estimated covariance matrix found by applying corrupted data $(\mathbb{X}, Y)_m^{n,k}$ in (1). Recall that, if the dimension of the e.d.r. subspace $\mathcal{K} \leq \min\{H, p\}$ were known, the SIR estimate of $\mathcal{B}$ would be defined by the subspace of $\mathbb{R}^p$ spanned by $\{\Sigma^{-1/2}\widehat{\eta}_{m,i}\}_{i=1}^{\mathcal{K}}$ (cf. Section 2.2) where $\{\widehat{\eta}_{m,i}\}_{i=1}^{\mathcal{K}}$ are the orthonormal eigenvectors of $\widehat{V}_{m,\mathrm{kn}}$ corresponding to the $\mathcal{K}$ largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$. If $\mathcal{K}$ were unknown, the same principle would apply using an estimate $\widehat{\mathcal{K}}$ of $\mathcal{K}$ determined by the number of large or dominating eigenvalues from $\widehat{V}_{m,\mathrm{kn}}$ (cf. Section 4). Regardless of any knowledge about $\mathcal{K}$, we can generally prescribe how the k-observation contamination scheme subsequently affects $\widehat{V}_{m,\mathrm{kn}}$ and subspace estimates based on its largest eigenvectors.

In the following, as $m \to \infty$, recall that the magnitude of contamination increases (i.e., defined by scaling $t_m \to \infty$) and that $\{\widetilde{\beta}_i\}_{i=1}^{k}$ denote directions of contamination from (3)–(4). If $\mathcal{A}$ denotes a generic vector subspace of $\mathbb{R}^p$, we let $P_{\mathcal{A}}$ denote the projection matrix for the subspace $\mathcal{A}$, where $P_{\mathcal{A}}$ and $\mathcal{A}$ uniquely define each other (e.g., if $\mathcal{A}$ is spanned by the independent columns of a matrix A then $P_{\mathcal{A}} = A(A^\top A)^{-1}A^\top$).

Theorem 1. *Suppose the k-slice contamination scheme from* Section 2.2, $1 \leq k \leq \min\{p, H\}$, *with known* $\mathrm{E}(\mathbb{X}) = \mu$ *and* $\mathrm{Cov}(\mathbb{X}) = \Sigma$. *Let* $n_{(1)} \leq \cdots \leq n_{(k)}$ *denote the ordering of the number* $\{n_h\}_{h=1}^k$ *of observations in the first k slices. From the SIR-dimension reduction procedure applied to the contaminated sample* $(\mathbb{X}, Y)_m^{n,k}$, *let* $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ *denote the p ordered eigenvalues of* $\widehat{V}_{m,\mathrm{kn}} \in \mathbb{R}^{p\times p}$ *and corresponding orthonormal eigenvectors* $\widehat{\eta}_{m,i}$, $i = 1, \ldots, p$. *As* $m \to \infty$,

(a) it holds that

$$\lim_{m\to\infty} \frac{\widehat{\lambda}_{m,h}}{t_m^2} = \begin{cases} \dfrac{1}{n_{(h)}n} & \textit{for} \quad h = 1, \ldots, k, \\ 0 & \textit{otherwise.} \end{cases}$$

(b) there exists a sequence Q_m of orthogonal $k \times k$ matrices such that

$$\lim_{m\to\infty} \left[\widehat{\eta}_{m,1} \cdots \widehat{\eta}_{m,k}\right] Q_m = \left[\Sigma^{-1/2}\widetilde{\beta}_1 \cdots \Sigma^{-1/2}\widetilde{\beta}_k\right].$$

(c) if $\widehat{\mathcal{N}}_m$ *and* $\Sigma^{-1/2}\widetilde{\mathcal{B}}$ *denote subspaces of* $\mathbb{R}^p$ *spanned by* $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ *and* $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^k$, *respectively, then,*

$$\lim_{m\to\infty} P_{\widehat{\mathcal{N}}_m} = P_{\Sigma^{-1/2}\widetilde{\mathcal{B}}}.$$

(d) if $\Sigma^{-1/2}\widehat{\mathcal{N}}_m$ *and* $\Sigma^{-1}\widetilde{\mathcal{B}}$ *denote the subspaces of* $\mathbb{R}^p$ *spanned by the estimated e.d.r. directions* $\{\widehat{\beta}_{m,h} = \Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^k$ *and the directions* $\{\Sigma^{-1}\widetilde{\beta}_h\}_{h=1}^k$ *of contamination, respectively, then*

$$\lim_{m\to\infty} P_{\Sigma^{-1/2}\widehat{\mathcal{N}}_m} = P_{\Sigma^{-1}\widetilde{\mathcal{B}}}.$$

(e) for a given $h \in \{1, \ldots, k\}$, *if* $\Sigma^{-1/2}\widehat{\mathcal{N}}_{m,h}$ *and* $\Sigma^{-1}\widetilde{\mathcal{B}}_h$ *denote subspaces of* $\mathbb{R}^p$ *spanned by* $\{\widehat{\beta}_{m,j} = \Sigma^{-1/2}\widehat{\eta}_{m,j} : 1 \leq j \leq k,\ n_j = n_{(h)}\}$ *and* $\{\Sigma^{-1}\widetilde{\beta}_j : 1 \leq j \leq k,\ n_j = n_{(h)}\}$, *respectively, then*

$$\lim_{m\to\infty} P_{\Sigma^{-1/2}\widehat{\mathcal{N}}_{m,h}} = P_{\Sigma^{-1}\widetilde{\mathcal{B}}_h}.$$

To comment on the significance of Theorem 1, part(a) shows that, upon contaminating k slices, the k largest eigenvalues of the contaminated matrix $\widehat{V}_{m,\mathrm{kn}}$ "explode" and grow at a rate faster than the remaining $p-k$ eigenvalues. Thus, as defined by the largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$, the "important" eigenvectors of $\widehat{V}_{m,\mathrm{kn}}$ for determining e.d.r. subspace estimation will necessarily correspond to $\{\widehat{\eta}_{m,h}\}_{h=1}^{k}$. However, by Theorem 1(c), the $\mathbb{R}^p$-subspace $\widehat{\mathcal{N}}_m$ spanned by these contaminated eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^{k}$ converges to the $\mathbb{R}^p$-subspace $\widetilde{\mathcal{B}}$ spanned by $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^{k}$ based on the k directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^{k}$ from (4) (i.e., the respective projection matrices converge). Or, alternatively viewed in matrix form in Theorem 1(b), the contaminated eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^{k}$, up to an orthogonal transformation, must converge to $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^{k}$. Consequently, in Theorem 1(d), when the regressor covariance Σ is known, the subspace $\Sigma^{-1/2}\widehat{\mathcal{N}}_m = \mathrm{span}\{\Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^{k}$ defined by the k (most important) contaminated eigenvectors must converge to the subspace $\Sigma^{-1}\widetilde{\mathcal{B}} = \mathrm{span}\{\Sigma^{-1}\widetilde{\beta}_h\}_{h=1}^{k}$ determined from the contaminated directions $\{\widetilde{\beta}_h\}_{h=1}^{k}$ in the first k slices. This finding is crucial because it reveals how the k-slice contamination scheme of Section 2.2 can directly influence the contaminated e.d.r. subspace estimate, which must necessarily contain all or part of $\Sigma^{-1}\widetilde{\mathcal{B}}$. Theorem 1(e), in fact, provides a type of generalization stating that, as the eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^{k}$ associated with the k largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$ have limits determined from the contaminated directions $\{\widetilde{\beta}_h\}_{h=1}^{k}$ used, the subspace spanned by certain sub-collections of estimated e.d.r. directions $\{\Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^{k}$ under contamination will converge to the subspace spanned by certain sub-collections of $\{\Sigma^{-1}\widetilde{\beta}_h\}_{h=1}^{k}$, depending the number of observations $\{n_h\}_{h=1}^{k}$ in the first k slices (these values define the limits of the k largest contaminated eigenvalues $\{\widehat{\lambda}_{m,h}\}_{h=1}^{k}$ in Theorem 1(a)). For example, if we assume $n_1 < \cdots < n_k$ for simplicity, then for *each* $i = 1, \ldots, k$ the subspace in $\mathbb{R}^p$ spanned by the first (i.e., most important) i estimated e.d.r. directions $\{\widehat{\beta}_{m,h} = \Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^{i}$ must converge to a subspace $\mathrm{span}\{\Sigma^{-1}\widetilde{\beta}_h\}_{h=1}^{i}$ determined by the first i directions of contamination. Therefore, the contamination scheme can control the dimension of the resulting estimated e.d.r. subspace under data contamination, as well as the subspace itself, by influencing how many and which contaminated eigenvalues are *large*.

Hence, as a direct consequence of Theorem 1, when the regressor covariance Σ is known and directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^{k}$ are *orthogonal* to the true e.d.r. directions $\{\beta_i\}_{i=1}^{\mathcal{K}}$ that span the e.d.r. subspace $\mathcal{B}$ (after a linear transformation of both contaminated and true directions by $\Sigma^{-1/2}$), the contaminated e.d.r. subspace estimate (or a crucial part of this) will be orthogonal to the true dimension reduction subspace. In this manner, contamination can effectively ruin attempts at dimension reduction by SIR. There admittedly remains a further technical matter in choosing a number k of contaminated points in order to make a contaminated estimate of $\mathcal{B}$ as distant as possible from this true subspace, which will be treated in Section 4 (e.g., depending on factors such as the number p of regressors, the dimension $\mathcal{K}$ of $\mathcal{B}$, and whether $\mathcal{K}$ is to be estimated). However, for the known Σ case, Theorem 1 provides the key ideas of how data contamination can directly distort e.d.r. subspace estimation, and numerical studies in Section 5 will further illustrate these findings in the case when contaminated directions are indeed orthogonal to true ones.

3.2. Results for Unknown Regressor Covariance. We next consider the effects of data contamination on SIR-based dimension reduction when the regressor covariance structure Σ is unknown. The overall influence of data contamination here differs substantially from the known covariance Σ case of Section 3.1 and, in a sense, is also more immediately damaging to the e.d.r. subspace estimation. This is because any subspace estimate will directly turn out to be orthogonal to *any* directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination used in the k-slice contamination scheme (Section 2.2). Hence, data contamination as defined in (3) in a direction $\widetilde{\beta}_h$ automatically forces an e.d.r. subspace estimate to lose that direction. To explain, note that the unknown mean and covariance structure of the regressors, μ and Σ, are replaced with data-based estimates in the SIR procedure (cf. Section 2.2). Recall that the data contamination involves (without loss of generality) replacing the first observation, denoted as $(\mathbf{x}_{1_h}, y_{1_h})$, $h = 1, \ldots, k$, in the first k slices with contaminated points $\{\widetilde{\mathbf{x}}_{m,1_h}, y_{1_h}\}$, $\widetilde{\mathbf{x}}_{m,1_h} = t_m\widetilde{\beta}_h + \widetilde{v}_h$ in (3). To estimate Σ based on $(\mathbb{X}, Y)_m^{n,k}$ based on a k-slice contaminated sample, we compute the sample covariance matrix, $\widehat{\Sigma}_m$, of the regressor observations, which can be algebraically rewritten as

$$\widehat{\Sigma}_m = S + \frac{n-k}{n^2} t_m^2 \sum_{h=1}^{k} \beta_{m,h}\beta_{m,h}^\top + \frac{k}{n^2} t_m^2 \sum_{h=1}^{k} (\beta_{m,h} - \overline{\beta}_m)(\beta_{m,h} - \overline{\beta}_m)^\top \quad (6)$$

where

$$S = \frac{1}{n} \sum_{\substack{i=1 \\ i \notin \{1_h : h=1,\ldots,k\}}}^{n} (\mathbf{x}_i - \overline{\mathbf{x}}^*)(\mathbf{x}_i - \overline{\mathbf{x}}^*)^\top, \quad \overline{\mathbf{x}}^* = \frac{1}{n-k} \sum_{\substack{i=1 \\ i \notin \{1_h : h=1,\ldots,k\}}}^{n} \mathbf{x}_i$$

and $\overline{\beta}_m = k^{-1}\sum_{h=1}^k \beta_{m,h}$ for $\beta_{m,h} = \widetilde{\beta}_h + \frac{\widetilde{v}_h - \overline{\mathbf{x}}^*}{t_m}$ for $h = 1, \ldots, k$. In (6), the sample covariance matrix $\widehat{\Sigma}_m$ is written into sums over uncontaminated (i.e., S) and contaminated observations. The dimension reduction subspace is now determined by the eigenvectors of a matrix $\widetilde{V}_{m,\text{un.kn}} = \widehat{\Sigma}_m^{-1} A_m \in \mathbb{R}^{p\times p}$, computed as in (2) from the contaminated data $(\mathbb{X}, Y)_m^{n,k}$, which has the form of $\widehat{\Sigma}_m^{-1}$ multiplied by another matrix, say, A_m. Because of this, understanding how the matrices $\widehat{\Sigma}_m^{-1}$ and $\widetilde{V}_{m,\text{un.kn}}$ behave in the presence of data contamination is crucial, which the next result addresses. In the following, let $\widetilde{\mathcal{B}} \subset \mathbb{R}^p$ denote the subspace spanned by the contaminated directions $\{\widetilde{\beta}_h\}_{h=1}^k$ and define the subspace $S^{-1/2}\widetilde{\mathcal{B}} = \text{span}\{S^{-1/2}\widetilde{\beta}_h\}_{h=1}^k \subset \mathbb{R}^p$, assuming the matrix S is positive definite in (6). Let I_p denote the $p \times p$ identity matrix, and write $\overline{\mathbf{x}}_h^* = \sum_{i\neq 1_h : y_i \in \mathbb{I}_h} \mathbf{x}_i/(n_h - 1)$ to denote the sample mean of the uncontaminated observations in the hth slice, $h = 1, \ldots, k$.

Theorem 2. *Under the k-slice contamination scheme of* Section 2.2*,* $1 \leq k \leq \min\{p, H\}$*, with unknown* $\mathrm{E}(\mathbb{X}) = \mu$ *and* $\mathrm{Cov}(\mathbb{X}) = \Sigma$*, suppose a contaminated sample* $(\mathbb{X}, Y)_m^{n,k}$ *yields an estimate* $\widehat{\Sigma}_m$ *of* Σ *as well as the matrix* $\widetilde{V}_{m,\text{un.kn}}$ *from* (2)*. Then, as* $m \to \infty$*,*

$$\lim_{m\to\infty} \widehat{\Sigma}_m^{-1} = S^{-1/2}\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}}\right) S^{-1/2}, \tag{7}$$

$$\lim_{m\to\infty} \frac{1}{t_m}\widetilde{V}_{m,\text{un.kn}} = S^{-1/2}\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}}\right) S^{-1/2} \sum_{h=1}^{k} \frac{n_h - 1}{n_h n}(\overline{\mathbf{x}}_h^* - \overline{\mathbf{x}}^*)\widetilde{\beta}_h^\top. \tag{8}$$

In this unknown regressor covariance Σ case, the behavior of $\widehat{\Sigma}_m^{-1}$, estimated from $(\mathbb{X}, Y)_m^{n,k}$, is again focal because of its appearance in the contaminated matrix version $\widetilde{V}_{m,\text{un.kn}} = \widehat{\Sigma}_m^{-1} A_m$ of (2). As the magnitude of contamination increases ($t_m \to \infty$ as $m \to \infty$), the positive definite inverse matrix $\widehat{\Sigma}_m^{-1}$ actually degenerates to a singular matrix in (7). But, more importantly, the null space of $\lim_{m\to\infty} \widehat{\Sigma}_m^{-1}$ is precisely the space spanned by the k directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination from (5). That is, the orthogonal complement of the column space of $S^{-1/2}\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}}\right) S^{-1/2} = \lim_{m\to\infty} \widehat{\Sigma}_m^{-1}$ is the $\mathbb{R}^p$-space $\widetilde{\mathcal{B}} = \text{span}(\{\widetilde{\beta}_h\}_{h=1}^k)$. The implications for dimension reduction are then severe. For example, if $\widehat{\beta}_m$ represents an estimated e.d.r. direction arising from the SIR procedure applied to contaminated data $(\mathbb{X}, Y)_m^{n,k}$, then $\widehat{\beta}_m$ is an eigenvector of $\widetilde{V}_{m,\text{un.kn}}$, or equivalently or $\widetilde{V}_{m,\text{un.kn}}/t_m$ in (8), and has a form

$$\widehat{\Sigma}_m^{-1} c_m/t_m = \widetilde{V}_{m,\text{un.kn}}\widehat{\beta}_m/t_m = \widehat{\lambda}_m\widehat{\beta}_m/t_m,$$

for some $c_m \in \mathbb{R}^p$ and $\widehat{\lambda}_m \in \mathbb{R}$. But, an estimated e.d.r. direction $\widehat{\beta}_m$ must belong to the column space of $\widehat{\Sigma}_m^{-1}$ (having the limit in (7)) and thereby become orthogonal to the subspace $\widetilde{\mathcal{B}} = \mathrm{span}(\{\widetilde{\beta}_h\}_{h=1}^k)$ as the contamination magnitude increases $m \to \infty$. In other words, the limit (as $m \to \infty$) of an estimated e.d.r. direction under contamination (i.e., an eigenvector of $\widetilde{V}_{m,\mathrm{un.kn}}/t_m$) must be in the column space of $S^{-1/2}\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}}\right)S^{-1/2}$ by (8) and hence *orthogonal to all directions of contamination* $\{\widetilde{\beta}_h\}_{h=1}^k$. To summarize, when Σ is unknown, a contaminated estimate of the dimension reduction subspace will lose (or be orthogonal to) *any* directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination. Hence, as a consequence of Theorem 2, when the regressor covariance Σ is unknown and the directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination span all or part of the true dimension reduction subspace $\mathcal{B} = \mathrm{span}(\{\beta_i\}_{i=1}^{\mathcal{K}})$, then the contaminated subspace estimate will necessarily be orthogonal to $\mathrm{span}(\{\widetilde{\beta}_h\}_{h=1}^k) \subset \mathcal{B}$ and thereby miss important information in dimension reduction. Section 5 provides numerical studies to concretely demonstrate these findings. Most importantly, in contrast to the known regressor covariance Σ case where harmful directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination are orthogonal to the true e.d.r. directions $\{\beta_i\}_{i=1}^{\mathcal{K}}$ (cf. Section 3.1), the most disruptive contaminated vectors $\{\widetilde{\beta}_h\}_{h=1}^k$ when Σ is unknown will lie in the *same* directions as the true e.d.r. directions $\{\beta_i\}_{i=1}^{\mathcal{K}}$.

4 Worst Case Data Contamination for Dimension Reduction and Breakdown Points

The results of Section 3 establish overall and general effects on how data contamination can disrupt dimension reduction and e.d.r. subspace estimation by SIR. To complement these findings, this section presents an investigation of *worst case behavior* of SIR under data contamination and corresponding breakdown points.

For SIR, and other dimension reduction procedures such as SAVE (Cook, 2000) or MAVE (Xia et al., 2002), a worst possible case of data contamination can be defined as producing an estimated subspace in $\mathbb{R}^p$, say $\widehat{\mathcal{B}}$, that is maximally distant from the true dimension reduction subspace $\mathcal{B}$ (e.g., $\widehat{\mathcal{B}}$ is as orthogonal as possible to $\mathcal{B}$). To formalize the concept of maximal distances between subspaces in $\mathbb{R}^p$, Section 4.1 briefly introduces a metric on subspaces. With this, worst case subspace estimation may be examined, but this process depends intricately on a combination of factors, such as the dimension p of the regressor space, the dimension $\mathcal{K}$ of the true e.d.r. subspace $\mathcal{B}$, and whether $\mathcal{K}$ is known or requires estimation. This study is further complicated by the issue that the most disruptive directions of contamination change between cases when the regressor covariance structure Σ is known or unknown. For clarity, Sections 4.2–4.3

separately treat worst case behavior in dimension reduction along these two cases.

4.1. A Metric on Subspaces and Failure in Dimension Reduction. Let $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ denote two vector subspaces of $\mathbb{R}^p$; if the columns of matrices $S = [s_1, \ldots, s_m]$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{m^*}]$ provide some orthonormal basis for $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively, having dimensions $m = \dim(\mathcal{S})$ and $m^* = \dim(\widetilde{\mathcal{S}})$, then $P_{\mathcal{S}} = SS^\top$ and $P_{\widetilde{\mathcal{S}}} = \widetilde{S}\widetilde{S}^\top$ are the projection matrices uniquely associated with these subspaces. To define a distance between $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$, we use a metric

$$\begin{aligned}\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) \equiv \|P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}\|_{\mathbb{F}} &= \left[\mathrm{tr}\left\{(P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}})(P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}})^\top\right\}\right]^{\frac{1}{2}} \\ &= [\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - 2\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}})]^{\frac{1}{2}}, \qquad (9)\end{aligned}$$

found by applying the Frobenius matrix norm, $\|\cdot\|_{\mathbb{F}}$, to the difference $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$ of projection matrices. $\mathbb{F}$ provides a generalization of a special subspace metric of Crone and Crosby (1995), but allowing the accommodation that the general subspaces of $\mathbb{R}^p$ may not have the same fixed dimension (unlike Crone and Crosby, 1995). The definition in (9) ensures that $\mathbb{F}$ constitutes a true metric on vector spaces (cf. Ch. II, Stewart and Sun, 1990, though these authors consider only $\mathbb{R}^p$ subspaces of the same dimension) and, while other matrix norms (e.g., Hölder norms) could similarly be used in place of $\|\cdot\|_{\mathbb{F}}$, these would generate the same topology as $\mathbb{F}$ in (9) on subspaces of $\mathbb{R}^p$ by the equivalence of such matrix norms. Further, $\mathbb{F}$ has some desirable properties in interpretation by incorporating the term $\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}})$, which may be viewed as a measure of the closeness of two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, or a numerical measure of what two subspaces have in common. It holds that $\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}}) \geq 0$ with this trace equaling zero if and only if the $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal, while this trace increases as the dimension of the intersection $\mathcal{S} \cap \widetilde{\mathcal{S}}$ increases. Krzanowski (1979) used this same trace quantity for assessing closeness in studies of principal components in multivariate analysis, though the trace $\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}})$ itself does not constitute a metric (e.g., by failing the triangle inequality).

To assess worst case behavior of SIR in dimension reduction, we may quantify how far off a contaminated subspace estimate $\widehat{\mathcal{B}}$ might be from the target e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$. For this, we distinguish between two cases depending on whether the dimension $\mathcal{K}$ of $\mathcal{B}$ is known, and would therefore be shared by an estimate $\widehat{\mathcal{B}}$ of $\mathcal{B}$, or is unknown and requires estimation.

The *maximal distance* between subspaces $\widehat{\mathcal{B}}$ and $\mathcal{B} \subset \mathbb{R}^p$ under (9) can be summarized as $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) =$

$$\begin{cases} \sqrt{2\big(\mathcal{K} - \max\{0, 2\mathcal{K} - p\}\big)} & \text{if } \dim(\mathcal{B}) = \mathcal{K} = \text{is known,} \\ \sqrt{p} & \text{otherwise } (\mathcal{K} \text{ is unknown/estimated).} \end{cases} \tag{10}$$

When $\mathcal{K}$ is known, an estimated subspace $\widehat{\mathcal{B}}$ is maximally distant from the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$ under $\mathbb{F}$ if their intersection $\mathcal{I} = \mathcal{S} \cap \widetilde{\mathcal{S}}$ is of the smallest possible dimension $\dim(\mathcal{I}) = \max\{0, 2\mathcal{K} - p\}$ and non-intersecting subspace portions $\mathcal{B} \cap \mathcal{I}^\perp, \widehat{\mathcal{B}} \cap \mathcal{I}^\perp$ are orthogonal. When $\mathcal{K}$ is unknown, a maximally distant estimate $\widehat{\mathcal{B}}$ is geometrically linked to estimation of the orthogonal complement $\widehat{\mathcal{B}} = \mathcal{B}^\perp$ of $\mathcal{B} \subset \mathbb{R}^p$, so that $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) = \sqrt{p}$ assumes the maximal value possible for the metric $\mathbb{F}$. In any event, whether $\mathcal{K}$ is known or unknown, worst case subspace estimation under data contamination is connected to estimation of orthogonal components of the true subspace $\mathcal{B}$, corresponding to the metric $\mathbb{F}$ assuming maximal distances between subspaces.

Based on the discussion above, we define the finite sample breakdown point for SIR.

Definition 4.1. ***Finite sample breakdown point.*** *Let $(\mathbb{X}, Y)^{n,k}$ denote a contaminated sample found by replacing $1 \le k \le n$ data points in a data set $(\mathbb{X}, Y)^n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ with arbitrary values $(\widetilde{\mathbf{x}}_{i_j}, \widetilde{y}_{i_j})_{j=1}^k$. Let $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_k \subset \mathbb{R}^p$ denote subspace estimates based on SIR applied to $(\mathbb{X}, Y)^n$ and $(\mathbb{X}, Y)^{n,k}$, respectively, where $\widehat{\mathcal{B}}_k$ has rank $1 \le \mathcal{K}_1 < p$ (say). Then, if the subspace rank $\mathcal{K}$ is known (not estimated by SIR so that $\mathcal{K}_1 = \mathcal{K}$), the* ***finite sample breakdown point*** *of SIR is defined as*

$$\epsilon_{fsbp,\mathcal{K}}((\mathbb{X}, Y)^n, \mathbb{F}) =$$

$$\min\left\{\frac{k}{n} : 1 \le k \le n, \sup_{(\mathbb{X},Y)^{n,k}} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_k) = \sqrt{2\big(\mathcal{K}_1 - \max\{0, 2\mathcal{K}_1 - p\}\big)}\right\}$$

under the metric $\mathbb{F}$ for the data constellation $(X, Y)^n$. Otherwise, the finite sample breakdown point is defined as

$$\epsilon_{fsbp,\mathcal{K}}((\mathbb{X}, Y)^n, \mathbb{F}) \quad = \quad \min\left\{\frac{k}{n} \, : \, 1 \le k \le n, \sup_{(X,Y)^{n,k}} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_k) = \sqrt{p}\right\}.$$

The value $\epsilon_{fsbp,\mathcal{K}}$ represents the percentage of contamination in a data set $(\mathbb{X},Y)^n$ necessary to cause a dimension reduction procedure to *breakdown* with a maximal distance (10) between the subspace estimate under contamination and the estimate produced from an original, uncontaminated sample. Note that for $\mathcal{K}=0$ or $\mathcal{K}=p$, the finite sample breakdown point definition above is not applicable nor meaningful in the $\mathcal{K}$ known case.

Remark 1. Our formulation (10) of a worst case subspace estimate agrees with that discussed by Becker (2001). However, we use a subspace metric $\mathbb{F}$ to quantify this, while Becker (2001) considered canonical correlations or principle angles between individual directions spanning subspaces (e.g., judging two subspaces as distant if their smallest canonical correlation is zero). While useful, canonical correlations do not directly evaluate distance between two entire subspaces (e.g., their computation is limited by the dimension of the smallest subspace). Relatedly, Hilker (1997) formulated "breakdown" of the SIR procedure by piecewise considering the breakdown properties of the individual location and scatter functionals found in the steps of SIR. In using a metric $\mathbb{F}$ to formulate worst case performance under contamination, we are more closely following the philosophy of Stromberg and Ruppert (1992) to look at "the performance of the procedure as a whole" in subspace estimation.

4.2. Results with Known Regressor Covariance. For the case that the regressor covariance $\Sigma \in \mathbb{R}^{p\times p}$ is known, we next provide results showing that, under the k-slice contamination scheme (Section 2.2), a contaminated data sequence $(\mathbb{X},Y)_m^{n,k}$ can be found so that the resulting contaminated subspace estimate $\widehat{\mathcal{B}}_m$ is worst case or maximally distant from the true e.d.r. subspace $\mathcal{B}\subset\mathbb{R}^p$ as in (10). Such results rely heavily on Theorem 1 and the findings from Section 3.1 where the most damaging directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination (used in formulating contaminated samples (3)) should be orthogonal to the true e.d.r. directions $\{\beta_i\}_{i=1}^{\mathcal{K}}$ spanning $\mathcal{B}$. The number k of contaminated data points needed for a worst case subspace estimate depends on whether $\mathcal{K}=\dim(\mathcal{B})$ is known.

When $\mathcal{K}$ is known, we formulate a data sample $(\mathbb{X},Y)_m^{n,k}$ with $k=\min\{\mathcal{K},p-\mathcal{K}\}$ contaminated points based on k directions of contamination essentially orthogonal to $\mathcal{B}$, as listed in Table 1; this suffices to produce worst case estimation (10) in this case. On the other hand, when $\mathcal{K}$ is unknown, we use $k=p-\mathcal{K}$ points with similarly contaminated directions; see Table 1. In this unknown $\mathcal{K}$ case, the contaminated sample $(\mathbb{X},Y)_m^{n,k}$ produces an estimate $\widehat{\mathcal{K}}_m$ by the testing criterion suggested in Li (1991), which is used to determine the dimension $\dim(\widehat{\mathcal{B}}_m)$ of the contaminated subspace estimate. This estimate $\widehat{\mathcal{K}}_m$ is essentially given by assessing the size, or rating the importance, of the largest eigenvalues $\widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ from $\widehat{V}_{m,\mathrm{kn}}$, the contaminated version of the covariance estimate from (2) (recall that if $\{\widehat{\eta}_{m,h}\}_{h=1}^{\widehat{\mathcal{K}}_m}$ are the orthonormal eigenvectors corresponding to the largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$, then the subspace estimate is $\widehat{\mathcal{B}}_m = \mathrm{span}(\{\Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^{\widehat{\mathcal{K}}_m})$ when Σ is known). Describing the effect of contamination on the estimate $\widehat{\mathcal{K}}_m$ is generally difficult, because the magnitude of all p eigenvalues $\{\widehat{\lambda}_{m,i}\}_{i=1}^{p}$ needs to be tracked and the test of dimension itself requires a critical value. To control the Type I error in Li (1991)'s testing procedure, one should usually choose a small significance level α, in accordance with the condition of Theorem 3. We now present the worst case estimation results with known regressor covariance.

Theorem 3. *Suppose the k-slice contamination scheme as described in* Table 1*, with known* $\mathrm{E}(\mathbb{X}) = \mu$ *and* $\mathrm{Cov}(\mathbb{X}) = \Sigma$*. From the SIR-dimension reduction procedure applied to the contaminated sample* $(\mathbb{X}, Y)_m^{n,k}$*, let* $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ *denote the p ordered eigenvalues of* $\widehat{V}_{m,\mathrm{kn}} \in \mathbb{R}^{p\times p}$ *and let* $\widehat{\mathcal{B}}_m$ *denote the resulting subspace estimate of $\mathcal{B} \subset \mathbb{R}^p$, having dimension $\mathcal{K}$ when* $\mathcal{K} = \dim(\mathcal{B})$ *is known and dimension $\widehat{\mathcal{K}}_m$ otherwise. Then, as $m \to \infty$,*

(a) when $\mathcal{K}$ is known,

$$\lim_{m\to\infty} \mathbb{F}(\widehat{\mathcal{B}}_m, \mathcal{B}) = \sqrt{2\big(\mathcal{K} - \max\{0, 2\mathcal{K} - p\}\big)}.$$

(b) when $\mathcal{K} < p$ is unknown,

$$\lim_{m\to\infty} \min_{h=1,\ldots,\mathcal{K}-p} \widehat{\lambda}_{m,h} = \infty, \qquad \limsup_{m\to\infty} \max_{h=\mathcal{K}-p+1,\ldots,p} \widehat{\lambda}_{m,h} \leq C$$

holds for some $C > 0$*, and there exists an* $\alpha^* \in (0,1)$ *where, if the testing procedure with significance level* $\alpha \in (0, \alpha^*]$ *is used to estimate the dimension* $\widehat{\mathcal{K}}_m$*,*

$$\lim_{m\to\infty} \widehat{\mathcal{K}}_m = p - \mathcal{K}, \qquad \lim_{m\to\infty} \mathbb{F}(\widehat{\mathcal{B}}_m, \mathcal{B}) = \sqrt{p}.$$

Hence, by Theorem 3, as the magnitude of contamination grows ($m \to \infty$), contamination can produce subspace estimates which are maximally distant (i.e., orthogonal) from the true e.d.r. subspace $\mathcal{B}$ and satisfy (10), both in the $\dim(\mathcal{B}) = \mathcal{K}$ known or unknown scenarios. In the more complex situation of $\mathcal{K}$ unknown, Theorem 3(b) shows that the contamination scheme, involving a replacement of $k = p - \mathcal{K}$ observations, forces the first $p - \mathcal{K}$ eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$ to become arbitrarily large while the last $\mathcal{K}$ eigenvalues are bounded and therefore relatively much smaller. In which case, it is natural that data contamination leads to a dimension estimate $\widehat{\mathcal{K}}_m \approx p - \mathcal{K}$ and further, with contaminated directions as in Table 1, the contaminated subspace estimate is then nearly equal to the orthogonal complement of the true e.d.r. subspace $\widehat{\mathcal{B}}_m \approx \mathcal{B}^{\perp}$. Therefore, in Theorem 3(b), the distance between $\widehat{\mathcal{B}}_m$ and $\mathcal{B}$ attains the largest possible value, $\sqrt{p}$, under the metric $\mathbb{F}$ for *any* two $\mathbb{R}^p$ subspaces. Such a distance is generally not achievable in the case where $\mathcal{K}$ is known and fixed (cf. Theorem 3(a)).

In accordance with Definition 4.1, we can also set an upper bound on the finite sample breakdown point of SIR based on a known covariance structure Σ, μ.

COROLLARY 4.1. *Suppose SIR seeks to estimate a $\mathcal{K}$-dimensional subspace of $\mathbb{R}^p$, $1 \le \mathcal{K} < p$, based on H data slices with a size n data sample $(\mathbb{X}, Y)^n$ and known values of $\mathrm{E}(\mathbb{X}) = \mu$, $\mathrm{Cov}(\mathbb{X}) = \Sigma$. Suppose the uncontaminated sample $(\mathbb{X}, Y)^n$ produces a subspace estimate $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ having rank $1 \le \mathcal{K}_1 < p$. Then, the finite sample breakdown point of SIR, under Definition 4.1, satisfies*

$$\epsilon_{fsbp,\mathcal{K}}((\mathbb{X}, Y)^n, \mathbb{F}) \;\le\; \frac{\min\{\mathcal{K}_1, p - \mathcal{K}_1\}}{n},$$

in both cases where the target $\mathcal{K}$ is known ($\mathcal{K} = \mathcal{K}_1$) or unknown.

For a proof of Corollary 4.1 we refer to the supplement Section 10.3.

4.3. Results with Unknown Regressor Covariance. When the regressor covariance $\Sigma \in \mathbb{R}^{p\times p}$ is unknown, we again find a contaminated data sequence $(\mathbb{X}, Y)_m^{n,k}$ with the resulting contaminated subspace estimate $\widehat{\mathcal{B}}_m$. However, based on the results of Section 3.2 (Theorem 2), the most harmful directions $\{\widetilde{\beta}_h\}_{h=1}^k$ of contamination in corrupted samples (c.f. (3)) should now be in the same direction as the true e.d.r. directions $\{\beta_i\}_{i=1}^{\mathcal{K}}$ spanning $\mathcal{B}$ (i.e., the estimate $\widehat{\mathcal{B}}_m$ will be orthogonal to any contaminated directions $\{\widetilde{\beta}_h\}_{h=1}^k$ used as a consequence of Theorem 2).

Table 1: For worst case studies in estimating true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$ ($\dim(\mathcal{B}) = \mathcal{K}$), the number and direction of contaminated data points in the k-slice contamination from Section 2.2

Σ, μ	number, k, when		directions of contamination, $\{\widetilde{\beta}_h\}_{h=1}^k$
	$\mathcal{K}$ known	$\mathcal{K}$ unknown	
Known	$\min\{\mathcal{K}, p-\mathcal{K}\}$	$p - \mathcal{K}$	as $\{\Sigma^{1/2}\overline{\beta}\}_{h=1}^k$ for orthonormal $\{\overline{\beta}\}_{h=1}^k \subset \{\Sigma^{1/2}v : v \in \mathcal{B}\}$
Unknown	$\min\{\mathcal{K}, p-\mathcal{K}\}$	$\mathcal{K}$	as independent vectors in $\mathcal{B}$

When $\mathcal{K}$ is known, we may produce a worst case subspace estimate $\widehat{\mathcal{B}}_m$ from a sample $(\mathbb{X}, Y)_m^{n,k}$ with $k = \min\{\mathcal{K}, p - \mathcal{K}\}$ contaminated points, involving k independent directions of contamination lying within the true e.d.r. subspace $\mathcal{B}$; see also the Table 1 summary. For $\mathcal{K}$ unknown, we use $k = \mathcal{K}$ points with similarly contaminated directions (i.e., taken from $\mathcal{B}$) to produce an estimate $\widehat{\mathcal{K}}_m$ (based on testing as in Section 4.2) and a contaminated subspace estimate $\widehat{\mathcal{B}}_m = \operatorname{span}(\{\widehat{\beta}_{m,h}\}_{h=1}^{\widehat{\mathcal{K}}_m})$ from $(\mathbb{X}, Y)_m^{n,k}$. However, in contrast to the case of Σ known and $\mathcal{K}$ unknown (Section 4.2), it now becomes intractable to formulate results on the size of the contaminated eigenvalues $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ of the matrix $\widetilde{V}_{m,\text{un.kn}}$ when Σ is *unknown*, in order to concretely determine the dimension $\dim(\widehat{\mathcal{B}}_m) = \widehat{\mathcal{K}}_m$, or how many eigenvectors $\{\widehat{\beta}_{m,h}\}_{h=1}^{p}$ of $\widetilde{V}_{m,\text{un.kn}}$ are needed to define the contaminated subspace estimate $\widehat{\mathcal{B}}_m$. In Section 3.2, we developed some strong results on the behavior of *eigenvectors* of $\widetilde{V}_{m,\text{un.kn}}$ (i.e., the corrupted e.d.r. directions) under contamination when Σ is unknown, but none of these results indicate the behavior of the *eigenvalues* of $\widetilde{V}_{m,\text{un.kn}}$ under contamination. In fact, the simulation study of Section 5 will show that, under various types of contamination as considered here, contaminated eigenvalues of $\widetilde{V}_{m,\text{un.kn}}$ may differ largely or very slightly from the eigenvalues of an uncontaminated matrix version $\widetilde{V}_{\text{un.kn}}$ from (2). Consequently, it is very difficult to control the estimated dimension $\widehat{\mathcal{K}}_m$ of the subspace estimate $\widehat{\mathcal{B}}_m$ under contamination.

The most complete statement about the worst case effect of contamination when Σ and $\mathcal{K}$ are both unknown can be summarized as follows. With the $k = \mathcal{K}$-point contamination mentioned above, a sequence of contaminated $\mathbb{R}^p$-subspace estimates $\widehat{\mathcal{B}}_m$ is produced that are orthogonal to the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$; in other words, framed in terms of projection matrices, $P_{\widehat{\mathcal{B}}_m} P_{\mathcal{B}} \approx 0_{p \times p}$ holds. Hence, one can ensure that the contaminated estimate is a subset of the orthogonal complement of $\mathcal{B} \subset \mathbb{R}^p$ (e.g., wrong information), but we are not able to guarantee that the dimension $\widehat{\mathcal{K}}_m$ of $\widehat{\mathcal{B}}_m$ is $p - \mathcal{K}$ (e.g., that $\widehat{\mathcal{B}}_m$ equals the full orthogonal complement of $\mathcal{B}$). Nevertheless, contamination can still drastically alter the subspace estimate by forcing estimation to be orthogonal to the target subspace $\mathcal{B}$.

We summarize the discussion above in the following formal statement. Additionally, Theorem 4 requires a further mild, though technical, condition for handling the Σ unknown case; this condition (a so-called data corruption pattern) is given in the supplement (Section 7) for brevity here.

Theorem 4. *Suppose the k-slice contamination scheme as in* Table 1, *along the data corruption pattern (in Supplement), with unknown* $\mathrm{E}(\mathbb{X}) = \mu$ *and* $\mathrm{Cov}(\mathbb{X}) = \Sigma$. *From the SIR-dimension reduction procedure applied to the* *contaminated sample* $(\mathbb{X},Y)^{n,k}_m$*, let* $\widehat{\mathcal{B}}_m$ *denote the resulting subspace estimate of* $\mathcal{B}\subset\mathbb{R}^p$*, having dimension* $\mathcal{K}$ *when* $\mathcal{K}=\dim(\mathcal{B})$ *is known and dimension* $\widehat{\mathcal{K}}_m$ *otherwise. Then, as* $m\to\infty$*,*

(a) when $\mathcal{K}$ *is known,*

$$\lim_{m\to\infty}\mathbb{F}(\widehat{\mathcal{B}}_m,\mathcal{B})=\sqrt{2\big(\mathcal{K}-\max\{0,2\mathcal{K}-p\}\big)}.$$

(b) when $\mathcal{K}$ *is unknown, the contaminated estimate* $\widehat{\mathcal{K}}_m=\dim(\widehat{\mathcal{B}}_m)$ *satisfies*

$$\limsup_{m\to\infty}\widehat{\mathcal{K}}_m\le p-\mathcal{K}$$

and $\widehat{\mathcal{B}}_m$ *is orthogonal to* $\mathcal{B}\subset\mathbb{R}^p$ *as* $m\to\infty$*, namely* $\lim_{m\to\infty}P_{\widehat{\mathcal{B}}_m}P_{\mathcal{B}}=0_{p\times p}$.

Finally, under Definition 4.1, we may also provide an upper bound on the finite sample breakdown point of a SIR-type procedure with an unknown covariance structure Σ.

Corollary 4.2. *Suppose SIR seeks to estimate a* $\mathcal{K}$*-dimensional subspace of* $\mathbb{R}^p$*,* $1\le\mathcal{K}<p$*, based on* H *data slices with a size* n *data sample* $(\mathbb{X},Y)^n$ *and unknown values of* $\mathrm{E}(\mathbb{X})=\mu$*,* $\mathrm{Cov}(\mathbb{X})=\Sigma$*. Suppose the uncontaminated sample* $(\mathbb{X},Y)^n$ *produces a subspace estimate* $\widehat{\mathcal{B}}\subset\mathbb{R}^p$ *having rank* $1\le\mathcal{K}_1<p$ *and that the data* $(\mathbb{X},Y)^n$ *satisfies the data corruption pattern (Supplement). Then, the finite sample breakdown point of SIR, under Definition 4.1, satisfies*

$$\epsilon_{fsbp,\mathcal{K}}((\mathbb{X},Y)^n,\mathbb{F})\ \le\ \frac{\mathcal{K}_1}{n},$$

in the case where the target dimension rank $\mathcal{K}$ *is known (*$\mathcal{K}=\mathcal{K}_1$*).*

When the covariance structure is known, a potentially lower bound on the finite sample breakdown point of SIR (cf. Cor 4.1) is possible compared to the unknown covariance case, intuitively because subspace estimation is more easily ruined by contamination in that case. When the dimension of the target subspace $\mathcal{K}$ is unknown in addition to the covariance structure, no breakdown point can be easily stated in Corollary 4 as, similarly to Theorem 4(b), it becomes difficult to generally control the size of a contaminated subspace estimate (though this can be made orthogonal to any target, e.g., $\widehat{\mathcal{B}}$ above). The next section examines the influence of data contamination on dimension reduction subspace estimation through simulation.

5 Simulation Study

We now numerically illustrate the main theoretical findings from Section 3 on the general effects of data contamination, which also served as the basis of the worst case studies in Section 4. Through simulation, we show that specific contamination patterns can yield subspace estimates that are orthogonal to the true e.d.r. subspace and thereby cause a loss of information in the regressor space dimension reduction.

Simulation Design Assume $\mathbb{X} = (X_1, X_2, X_3, X_4)^\top$ with $X_i \sim \mathcal{N}(0, 1)$, $i = 1, \ldots, 4$ so $\mathrm{E}(\mathbb{X}) = 0_4$ and $\mathrm{Cov}(\mathbb{X}) \equiv \Sigma = \mathrm{I}_4$. We consider two models which have been investigated previously in Hilker (1997) and Bond (1999), namely $Y = X_1 + X_2 + X_3 + X_4$ (Model 1) and $Y = X_1/(0.5 + \sqrt{|1.5 + X_2|})$ (Model 2), respectively. Similar models have also been used in Prendergast (2005, 2006). For simplicity, we do not include a further error term ε in these models. For Model 1, the true e.d.r. subspace has dimension $\mathcal{K} = 1$ spanned by $\beta_1 = (1, 1, 1, 1)^\top/\sqrt{4}$, whereas $\mathcal{K} = 2$ holds for Model 2 with two e.d.r. directions $\beta_1 = (1, 0, 0, 0)^\top$ and $\beta_2 = (0, 1, 0, 0)^\top$. Depending on the model and the knowledge of Σ, we explore various amounts of contamination by replacing either one, two or three observations in $(\mathbb{X}, Y)^n \in \mathbb{R}^4 \times \mathbb{R}$. Without loss of generality we replace the first $\mathbf{x}$-value, $\mathbf{x}_{1_h}$, by $\widetilde{\mathbf{x}}_{m,1_h} = t_m \widetilde{\beta}_h$, $\widetilde{\beta}_h \in \mathbb{R}^4$, in each of the first h slices, $h = 1, 2, 3$, as described in the contamination scheme of Section 2.2. Note that $\widetilde{\beta}_h$ denotes the direction of contamination and $t_m > 0$ is a scaling factor. We vary the magnitude of contamination as $t_m = 10^{2i}$ for $i = 0, 0.5, 1, 2, 3$. For clarity and brevity, we will present findings for Model 1 only in the case where Σ is known and findings for Model 2 in the Σ unknown case; full results, which are qualitatively similar, can be found in Genschel (2005). For Models 1 and 2, the contaminated directions $\widetilde{\beta}_h$, $h = 1, 2, 3$ are defined as follows depending on whether Σ is assumed to be known.

Σ is Known, Model 1: Effective data contamination involves orthonormal directions that are orthogonal to the true e.d.r. directions; see Table 1 and Theorems 1 and 3. For Model 1, the orthogonal complement of the true e.d.r. subspace $\mathrm{span}(\beta_1) \subset \mathbb{R}^4$ should theoretically be estimated if we replace one regressor observation in the first three slices with contaminated versions based on orthonormal contaminated directions $\{\widetilde{\beta}_h\}_{h=1}^3$ that are orthogonal to β_1; choices can be $\widetilde{\beta}_1 = (1, -1, 0, 0)^\top/\sqrt{2}$, $\widetilde{\beta}_2 = (0, 0, -1, 1)^\top/\sqrt{2}$ and $\widetilde{\beta}_3 = (1, 1, -1, -1)^\top/\sqrt{4}$ for example. Contamination of only one or two observations in these directions should result in estimation of only parts of the orthogonal complement.

Σ Unknown, Model 2: Contamination is used in the directions of the true e.d.r. directions; see also Table 1. Hence, by the results of Theorems 2 and 4, we replace either one or two observations in distinct slices using directions of contamination that span the true e.d.r. subspace for Model 2, i.e. $\widetilde{\beta}_h = \beta_h$, $h = 1, 2$. (In contrast, for Model 1, poor estimation would follow by our results when replacing one observation in the direction of $\beta_1 = (1, 1, 1, 1)^\top$.)

Performance Criteria. For a contaminated data sample, we let $\widehat{\beta}_i \equiv \widehat{\beta}_{m,i}$, $i = 1, \ldots, 4$, denote the estimated e.d.r. directions, where $\widehat{\beta}_i$ is associated with the ith largest contaminated eigenvalue $\widehat{\lambda}_i \equiv \widehat{\lambda}_{m,i}$, $i = 1, \ldots, 4$ (i.e., $\widehat{\lambda}_4 \leq \cdots \leq \widehat{\lambda}_1$) of either $\widehat{V}_{m,\text{kn}}$ (for known Σ) or $\widetilde{V}_{m,\text{un.kn}}$ (for unknown Σ); recall these latter matrices are the contaminated versions of the covariance estimates (1) and (2) used in SIR. The performance of estimation under contamination is evaluated through the vector product $\widehat{\beta}_i^\top \beta_1 = \cos(\widehat{\beta}_i^\top \angle \beta_1)$, $i = 1, \ldots, 4$, (additionally $\widehat{\beta}_i^\top \beta_2$ for Model 2), which is 0 when $\widehat{\beta}_i, \beta_1$ are orthogonal and approximately 1 or -1 when $\widehat{\beta}_i, \beta_1$ span the same direction. This vector product is used in place of the subspace metric (9) for these numerical studies to investigate the individual behavior of e.d.r. directions estimated under contamination, which perhaps captures finer detail. Because of orthonormality, one may evaluate the subspace metric (9) here from the reported results on individual directions (e.g., under Model 2, a trace correlation between between true and estimated subspaces follows by adding results comparing $\widehat{\beta}_i$ to β_j, $i, j = 1, 2$).

For each model and level of contamination, we conducted $M = 1000$ simulation runs generating data sets $(\mathbb{X}, Y)^n$ of sample size $n = 100$. The averages of $|\cos(\widehat{\beta}_i^\top \angle \beta_j)|$ for $i = 1, \ldots, 4$, $j = 1$ or 2, were computed along with the average values of ordered eigenvalues $\{\widehat{\lambda}_i\}_{i=1}^4$. For comparison, we also include these average values as computed from the *uncontaminated* data sets $(\mathbb{X}, Y)^n$ (i.e., prior to contamination as above). Results for Model 1 (Σ known) and Model 2 (Σ unknown) are presented in Sections 5.1 and 5.2, respectively.

5.1. Simulation results when Σ is known. Table 2 displays several results obtained for Model 1, beginning with a single contaminated observation $k = 1$ in the direction of $\widetilde{\beta}_1 = (1, -1, 0, 0)^\top/\sqrt{2}$, i.e., orthonormal to the true e.d.r. direction $\beta_1 = (1, 1, 1, 1)^\top$. A main result from Section 3 was

Table 2: Model 1, Σ known: Estimated eigenvalues and e.d.r. directions w.r.t. $\beta_1 = (1,1,1,1)^\top$ under contamination. One contaminated point in the direction of $\widetilde{\beta}_1 = (1,-1,0,0)^\top/\sqrt{2}$

Contamination	Estimated eigenvalues			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	0.89943299	0.14011246	0.06967781	0.02807165
$t_m = 10^1$	0.93754552	0.17523492	0.08204017	0.03292336
$t_m = 10^2$	9.35701843	0.62411524	0.10242710	0.03875058
$t_m = 10^4$	8.999768e+04	6.375259e-01	1.011727e-01	3.766357e-02
$t_m = 10^6$	9.000002e+08	6.362138e-01	1.006758e-01	3.855718e-02
No contamination	0.98671578	0.14187920	0.07006886	0.02914564
	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \beta_1)\rvert$
$t_m = 10^0$	0.979841	0.09456183	0.09123813	0.0904320
$t_m = 10^1$	0.9591784	0.1581040	0.1250104	0.1117938
$t_m = 10^2$	0.1750000	0.964122	0.1140435	0.1086621
$t_m = 10^4$	0.001672263	0.9797691	0.1204768	0.1043951
$t_m = 10^6$	1.666798e-05	0.9788298	0.1198380	0.1119880
No contamination	0.981705	0.09123825	0.0871689	0.08483905

that, when Σ is known, contamination of k slices should cause the k largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$ to “explode” and to grow infinitely large at a rate faster than the remaining $p-k$ eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$; see Theorem 1 for example. From Table 2, we see that the under contamination estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$ clearly support this finding. With $k = 1$ contamination point, not only is the first estimated eigenvalue affected by the contamination itself, as the magnitude t_m increases we have indeed that $\widehat{\lambda}_1$ grows infinitely large at a rate faster than the remaining eigenvalues $\widehat{\lambda}_2$, $\widehat{\lambda}_3$ and $\widehat{\lambda}_4$. The same effect can be observed in Tables 3 and 4 for cases with two or three contaminated observations, respectively, where the number $k = 2, 3$ of contaminated points exactly matches the number of eigenvalues exploding in each case (i.e., related to the number of estimated e.d.r. directions deemed as important under contamination).

In connection to the eigenvalue estimates, Tables 2–4 correspondingly display the behavior of estimated e.d.r. directions under data contamination. As more contaminated points are used, we find that the estimated directions $\widehat{\beta}_i$ under contamination, corresponding to exploding eigenvalues $\widehat{\lambda}_i$, are orthogonal to the true e.d.r. direction β_1. This behavior exactly matches our theoretical findings in Theorem 1. Namely, the space spanned by those estimated e.d.r. directions associated with the k largest eigenvalues of $\widehat{V}_{m,\mathrm{kn}}$ (under k-replacement contamination) converges to the subspace spanned by the k directions of contamination, a subspace which is indeed orthogonal to

Table 3: Model 1, Σ known: Estimated eigenvalues and e.d.r. directions w.r.t. $\beta_1 = (1,1,1,1)^\top$ under contamination. Two contaminated points in the directions of $\widetilde{\beta}_1 = (1,-1,0,0)^\top/\sqrt{2}$ and $\widetilde{\beta}_2 = (0,0,-1,1)^\top/\sqrt{2}$

Contamination	Estimated eigenvalues			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	0.87849150	0.13586764	0.06707110	0.02734312
$t_m = 10^1$	0.92174541	0.21847900	0.10502371	0.04010948
$t_m = 10^2$	10.23292101	8.40767119	0.43498218	0.05851056
$t_m = 10^4$	9.999739e+04	8.000031e+04	4.515726e-01	5.869024e-02
$t_m = 10^6$	1.000000e+09	7.999998e+08	4.545222e-01	6.051161e-02
No contamination	0.97802389	0.14168882	0.06963360	0.02806297
	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_1)\rvert$
$t_m = 10^0$	0.9796738	0.0946602	0.09170605	0.08955545
$t_m = 10^1$	0.9468494	0.1469614	0.1685776	0.1377121
$t_m = 10^2$	0.06132069	0.2108020	0.9572554	0.1432630
$t_m = 10^4$	0.0003923886	0.002162707	0.985721	0.1317057
$t_m = 10^6$	3.972363e-06	2.177360e-05	0.9849557	0.138369
No contamination	0.980554	0.09614678	0.08870456	0.08595423

the true e.d.r. subspace span(β_1), providing numerical evidence that data contamination can influence the dimension and directions of the estimated reduction subspace.

5.2. Simulation Results When Σ is Unknown. Numerical summaries under Model 2 appear in Tables 5–8, where Tables 5 and 6 show results for estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$ and Tables 7 and 8 display corresponding results for estimated e.d.r. directions $\widehat{\beta}_1, \ldots, \widehat{\beta}_4$, where all tables describe situations with varying amounts of contamination.

With two points $k = 2$ of contamination (each in a true e.d.r. direction β_1, β_2) in Table 5, the largest eigenvalue $\widehat{\lambda}_1$ shows the greatest amount of change and appears to decrease most substantially relative to the uncontaminated versions of the eigenvalues. While the presence of contamination does indeed influence estimated eigenvalues, these eigenvalues do not eventually change as t_m increases so that the magnitude t_m of contamination has only a limited effect on estimated eigenvalues; this behavior when Σ is unknown differs substantially from the numerical findings in Section 5.1 for the known Σ case. However, these results agree with the theory established in that the size of eigenvalues under contamination are not so clearly controlled when Σ is unknown, in contrast to the Σ known case.

Table 6 shows the behavior of estimated eigenvalues when we contaminate only $k = 1$ observation $\mathbf{x}_{1_1}$, either contaminated in the direction of one true e.d.r. direction β_1 or in the direction of the other β_2. An interesting

Table 4: Model 1, Σ known: Estimated eigenvalues and e.d.r. directions w.r.t. $\beta_1 = (1,1,1,1)^\top$ under contamination. Three contaminated points in the directions of $\widetilde{\beta}_1 = (1,-1,0,0)^\top/\sqrt{2}$, $\widetilde{\beta}_2 = (0,0,-1,1)^\top/\sqrt{2}$ and $\widetilde{\beta}_3 = (1,1,-1,-1)^\top/\sqrt{4}$

Contamination	Estimated eigenvalues			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	0.85937553	0.13956426	0.06805860	0.02836590
$t_m = 10^1$	0.9229757	0.2560026	0.1349992	0.0569284
$t_m = 10^2$	10.706793	9.648394	7.447724	3.02329e-01
$t_m = 10^4$	1.000544e+05	9.994727e+04	6.999694e+04	3.216579e-01
$t_m = 10^6$	1.000005e+09	9.999947e+08	7.000002e+08	3.254403e-01
No contamination	0.98090816	0.14086985	0.07032647	0.02850672
	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_1)$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_1)\rvert$
$t_m = 10^0$	0.978474	0.1010335	0.09293968	0.09020317
$t_m = 10^1$	0.9451667	0.1352988	0.1440981	0.1852461
$t_m = 10^2$	0.05339378	0.0584024	0.2467157	0.9640377
$t_m = 10^4$	0.0004063888	0.0004063516	0.002578214	0.9999964
$t_m = 10^6$	4.145536e-06	3.837078e-06	2.576043e-05	1
No contamination	0.980692	0.09342744	0.09005521	0.08810476

dichotomy emerges on the effect of contamination. That is, contamination in the direction of β_1 results in a corruption of all estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$, which decrease in value compared to the case of no contamination and change in the largest contaminated eigenvalue is most dramatic (similarly to the $k = 2$ point contamination mentioned above). However, a contamination in the direction of β_2 in Table 6 causes very little change in $\widehat{\lambda}_1$, or the other estimated eigenvalues, compared to estimation with uncontaminated data. Hence, perhaps surprisingly, behavior of the estimated eigenvalues from SIR under contamination is not symmetric with respect to the directions of contamination when the regressor covariance Σ is unknown. Consequently, the influence of the directions of contamination in determining eigenvalue estimates is quite complicated in this case, which numerically

Table 5: Model 2, Σ unknown: Estimated eigenvalues under contamination in the directions of true e.d.r. directions $\beta_1 = (1,0,0,0)^\top$, $\beta_2 = (0,1,0,0)^\top$

Contamination	Estimated eigenvalues			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	0.85684405	0.17707342	0.08672674	0.03419912
$t_m = 10^1$	0.36959644	0.13307111	0.07107053	0.03049050
$t_m = 10^2$	0.16238775	0.09766174	0.05719913	0.02642980
$t_m = 10^4$	0.16132208	0.10475009	0.06587638	0.02935102
$t_m = 10^6$	0.16306890	0.10474233	0.06671313	0.03074495
No contamination	0.92968171	0.18887460	0.08852808	0.03577398

Table 6: Model 2, Σ unknown: Estimated eigenvalues under contamination in only one direction of the true e.d.r. directions $\beta_1 = (1,0,0,0)^\top$, $\beta_2 = (0,1,0,0)^\top$

Contamination	Estimated eigenvalues $\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
One contaminated point in the direction $\beta_1 = (1,0,0,0)^\top$ only				
$t_m = 10^0$	0.86775173	0.18095971	0.08666652	0.03475568
$t_m = 10^1$	0.36338986	0.16608091	0.08396544	0.03436439
$t_m = 10^2$	0.20920058	0.10947518	0.05699737	0.02395002
$t_m = 10^4$	0.21246738	0.11486716	0.06303169	0.02658654
$t_m = 10^6$	0.21185503	0.11654676	0.06345007	0.02741657
No contamination	0.92949550	0.18676763	0.08929240	0.03601315
One contaminated point in the direction $\beta_2 = (0,1,0,0)^\top$ only				
$t_m = 10^0$	0.90802671	0.17101349	0.08452063	0.03463706
$t_m = 10^1$	0.92057744	0.12817938	0.06336271	0.02574777
$t_m = 10^2$	0.93192260	0.12663974	0.06482562	0.02770467
$t_m = 10^4$	0.93175652	0.13156406	0.06773224	0.02981461
$t_m = 10^6$	0.93231577	0.12961365	0.06835596	0.03071698
No contamination	0.92919865	0.18293314	0.08698891	0.03665909

supports the difficulty in theoretically isolating the asymptotic behavior of eigenvalues under contamination when Σ is estimated compared to the setting where Σ is known (cf. compare Theorems 1–3 to Theorems 2–4).

Table 7 shows the results for the estimated e.d.r. directions under contamination corresponding to the eigenvalues found in Table 5. Without contamination, SIR appears to correctly estimate β_1 but is somewhat less successful in estimating β_2. The contamination of two observations $k = 2$ in the directions of β_1 and β_2, respectively, causes SIR to estimate all e.d.r. directions as orthogonal to the true ones, β_1 and β_2. That is, none of the estimated e.d.r. directions, regardless of the corresponding eigenvalues, are elements of the subspace spanned by β_1 and β_2. This aspect is in perfect agreement with the theoretical findings in Section 3.2.

Table 8 displays results for the estimated e.d.r directions obtained from using one contaminated observation $k = 1$ in the direction of $\beta_2 = (0,1,0,0)^\top$. Just as SIR was able to accurately estimate $\widehat{\lambda}_1$ under this form of contamination (Table 6), the procedure also correctly estimates β_1. However, as a consequence of contamination in the direction of β_2, we find that all contaminated estimates $\{\widehat{\beta}_i\}_{i=1}^4$ are orthogonal to the true e.d.r. direction β_2, which represents a direct loss of information due to this contamination.

In addition to confirming our theoretical results, the simulation study also helps to explain earlier findings in the literature. In a data example of Sheather and McKean (2001), their Fig. 12 shows two clear outliers, cases 11 and 157, in directions almost orthogonal to the one explaining the largest

Table 7: Model 2, Σ unknown: Behavior of estimated e.d.r. directions under contamination where true e.d.r. directions are $\beta_1 = (1, 0, 0, 0)^\top$, $\beta_2 = (0, 1, 0, 0)^\top$

Contamination	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_1)\rvert$
Behavior w.r.t. β_1 with two contaminated points in directions β_1, β_2				
$t_m = 10^0$	0.9973476	0.0972353	0.09455044	0.09119261
$t_m = 10^1$	0.8950354	0.1805444	0.1802694	0.1612251
$t_m = 10^2$	0.05231419	0.08107545	0.1117936	0.0875233
$t_m = 10^4$	0.0006141142	0.001902911	0.001151127	0.0004754775
$t_m = 10^6$	5.428481e-06	1.862893e-05	1.235388e-05	4.690284e-06
No contamination	0.9981598	0.08652421	0.08502864	0.0850214
Contamination	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_2)\rvert$
Behavior w.r.t. β_2 with two contaminated points in directions β_1, β_2				
$t_m = 10^0$	0.04509871	0.7150232	0.4017536	0.3328404
$t_m = 10^1$	0.3173347	0.3630845	0.4169581	0.3783862
$t_m = 10^2$	0.07831053	0.1022642	0.08702715	0.06199629
$t_m = 10^4$	0.0006027861	0.001912108	0.001109116	0.0004811929
$t_m = 10^6$	5.428294e-06	1.890622e-05	1.218958e-05	4.619279e-06
No contamination	0.07831053	0.1022642	0.08702715	0.06199629

amount of regressor variability (which should be reasonably close to the estimated e.d.r. direction) and the exclusion of both outliers produced little effect on $\widehat{\beta}_1$ (see Fig. 15 in Sheather and McKean, 2001). This numerical result is in agreement with our findings. When Σ is unknown, estimated e.d.r. directions will be orthogonal to the directions of strong contamination. Consequently, in the data example, contamination orthogonal to β_1, such as cases 11 and 157, is not expected then to have much impact on the resulting estimate $\widehat{\beta}_1$. Similarly, susceptibility of SIR may be anticipated when extreme observations happen to fall in a true e.d.r. direction, which indeed is plausible when dealing with regressor variables exhibiting heavy-tailed distributions as, for example, considered by Chiancone et al. (2016).

In summary, the simulation evidence strongly supports the theory established in Section 3 on the effects of data contamination in dimension reduction. When the regressor covariance Σ is known, the amount and directions of contamination can immediately control the size of contaminated eigenvalues (i.e., which estimated directions are relatively important) and induce wrong subspace estimates (i.e., the estimated e.d.r. directions associated with these large eigenvalues are orthogonal to the true reduction subspace). When Σ is unknown, the effect of contamination on estimated eigenvalues is not easily to quantify but, in a sense, this aspect is not so important in this case. With directions of contamination lying among the true

Table 8: Model 2, Σ unknown: Behavior of estimated e.d.r. directions under contamination where true e.d.r. directions are $\beta_1 = (1, 0, 0, 0)^\top$, $\beta_2 = (0, 1, 0, 0)^\top$

Contamination	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_1)\rvert$
Behavior w.r.t. β_1 with one contaminated point in direction β_2				
$t_m = 10^0$	0.997704	0.09540198	0.08763669	0.0851936
$t_m = 10^1$	0.9947148	0.1068113	0.1190944	0.1354513
$t_m = 10^2$	0.9988849	0.1162464	0.1884353	0.1964554
$t_m = 10^4$	0.9989754	0.128747	0.2022079	0.1719973
$t_m = 10^6$	0.9989905	0.1268841	0.2079789	0.1703580
No contamination	0.9980909	0.08365931	0.08509767	0.08603543
Contamination	Cosine of angle between e.d.r. directions			
	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_2)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_2)\rvert$
Behavior w.r.t. β_2 with one contaminated point in direction β_2				
$t_m = 10^0$	0.0409813	0.7044577	0.416999	0.3324463
$t_m = 10^1$	0.0903446	0.2760754	0.3938079	0.5353305
$t_m = 10^2$	0.01692782	0.04406396	0.08927996	0.09609494
$t_m = 10^4$	0.0001718877	0.000515331	0.001011452	0.0008247947
$t_m = 10^6$	1.723068e-06	5.011812e-06	1.008949e-05	8.171434e-06
No contamination	0.03353829	0.7515609	0.3812316	0.3143829

e.d.r. directions, any estimated e.d.r. directions under contamination will be orthogonal to true e.d.r. directions, proving an immediate destruction of information.

6 Conclusions and Recommendations

The aim of this paper was a detailed and formal investigation of the sensitivity of dimension reduction to data contamination, which has not been well understood in the literature. We focused on the dimension reduction procedure SIR (Li, 1991) to facilitate this study, but also because SIR has been considered in other contexts of data contamination (cf. Section 1) with mixed reports on the effects of contamination. We have shown that data contamination scenarios *can* indeed produce completely incorrect dimension reduction subspace estimates in SIR thus confirming results by Gather et al. (2002) for $\mathcal{K} = 1$ and extending the existing results to general $\mathcal{K} > 1$. General results indicated, in an overall manner (cf. Section 3), how the amount (i.e., number of contaminated data points), magnitude and directions of contamination influence subspace estimates and that the "right" directions of contamination will bend an estimated subspace away from the true or target reduction subspace. Hence, outlying observations may not, simply by being outliers, be detrimental to dimension reduction, but outlying observations in appropriate directions can seriously corrupt subspace estimation. This aspect seems to explain, and perhaps unify previous findings in the literature about the effects of data contamination on SIR. While data contamination can force important e.d.r. directions to become utterly lost in estimation, the type of data contamination that can cause SIR to yield an erroneous subspace estimate changes intricately depending on whether the regressor covariance structure is known or not. Our theoretical findings were followed by a simulation study in Section 5, which demonstrated and supported our established theory. Finite sample breakdown points were also formulated for quantifying the global robustness of SIR or other dimension reduction procedures.

Our study leads us to believe that conclusions based on the sensitivity of SIR to data contamination may not apply immediately to other methods for dimension reduction, such as SAVE (Cook, 2000), MAVE (Xia et al., 2002) or robustified versions of these, which rather deserve to be studied in their own right. Further research on data contamination in high dimensional settings is necessary to better understand the notion and effect of outliers as well as inliers.

7 Supplement

To save space, proofs of all results are provided in this supplement. Additionally, for establishing these results, the supplement contains additional technical lemmas concerning the asymptotic properties of several types of matrix perturbations, which may be of independent interest.

Acknowledgements. The author is grateful to two anonymous reviewers for providing thoughtful comments improving the manuscript. The author would also like to thank Ursula Gather for suggesting and supporting this research direction.

References

BECKER, C. (2001) *Robustness Concepts for Analyzing Structured and Complex Data Sets.* Habilitationsschrift, University of Dortmund.

BOND, J.C. (1999) *A Robust Approach to SIR.* PhD Thesis, University of California, Berkeley.

CHEN, X., ZHOU, C. and COOK, R.D. (2010) Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.* **38**, 3696–3723.

CHIANCONE, A., FORBES, F. and GIRARD, S. (2016) Student Sliced Inverse Regression. *Computational Statistics and Data Analysis, Elsevier, 2016*, doi: 10.1016/j.csda.2016.08.004. hal-01294982v3.

COOK, R.D. (2000) A method for dimension reduction and graphics in regression. *Commun. Statist.-Theory Meth.* **829**, 2109–2121.

COOK, R.D. and CRITCHLEY, F. (2000) Identifying regression outliers and mixtures graphically. *J. Amer. Statist. Assoc.* **95**, 81–794.

COOK, R.D. and NI, L. (2005) Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. *J. Am. Statist. Ass.* **100**, 410–428.

COOK, R.D. and NI, L. (2006) Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika* **93**, 65–74.

COOK, R.D. and YIN, X. (2000) Dimension reduction and visualization in discriminant analysis (with Discussion). *Aust. N. Z. J. Stat.* **43**(2), 147–199.

COOK, R.D. and WEISBERG, S. (1991) Comment on: Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86**, 328–332.

CRONE, L.J. and CROSBY, D.S. (1995) Statistical applications of a metric on subspaces to satellite meteorology. *Technometrics* **37**, 324–328.

CROUX, C. and HAESBROECK, G. (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika* **87**(3), 603–618.

CROUX, C. and RUIZ-GAZEN, A. (2005) High breakdown estimators for principal components: The projection-pursuit approach revisited. *J. Multivariate Anal.* **95**, 206–226.

DONG, Y., YU, Z. and ZHU, L. (2015) Robust inverse regression for dimension reduction. *J. Multivariate Anal.* **135**, 71–81.

FORZANI, L., COOK, R.D. and ROTHMAN, A.J. (2012) Estimating sufficient reductions of the predictions in abundant high-dimensional regressions. *Ann. Statist.* **40**(1), 353–384.

GATHER, U., HILKER, T. and BECKER, C. (2002) A note on outlier sensitivity of sliced inverse regression. *Statistics* **13**, 271–281.

GATHER, U., HILKER, T. and BECKER, C. (2001) A Robustified Version of Sliced Inverse Regression. In: FERNHOLZ, L.T., MORGENTHALER, S., STAHEL, W. (eds.) *Statistics in Genetics and in the Environmental Sciences, Proceedings of the Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics held in Ascona from May 23–28 1999, 147–157.*

GENSCHEL, U. (2005) *Robustness Concepts for Sliced Inverse Regression.* PhD Thesis, University of Dortmund.

GENSCHEL, U. (2017) *Supplement to on the influence of data contamination in dimension reduction.*

HILKER, T. (1997) *Robuste Verfahren zur Dimensionsreduktion in Regressionsverfahren mit unbekannter Linkfunktion.* PhD Thesis, University of Dortmund.

KRZANOWSKI, W.J. (1979) Between-groups comparison in principal components. *J. Amer. Statist. Assoc.* **74**, 703–707.

LI, K.-C. (1991) Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316–342.

LI, K.-C. (1992) On principal hessian directions for data visualization and dimension reduction: Another application of Stein's lemma. *J. Amer. Statist. Assoc.* **87**, 1025–1039.

LI, L. (2007) Sparse sufficient dimension reduction. *Biometrika* **94**, 603–613.

LI, G. and CHEN, Z. (1985) Projection-pursuit approach to robust dispersion matrices and principal components: Primary theory and Monte Carlo. *J. Am. Statist. Ass.* **80**, 759–766.

LI, B. and DONG, Y. (2009) Dimension reduction for non-elliptically distributed predictors. *Ann. Statist.* **37**(3), 1272–1298.

LI, L., LI, B. and ZHU, L.-X. (2010) Groupwise dimension reduction. *J. Am. Statist. Ass.* **105**, 1188–1201.

ma, y. and zhu, l. (2012) A semiparametric approach to dimension reduction. *J. Am. Statist. Ass.* **107**, 168–179.

ma, y. and zhu, l. (2013) Efficient estimation in sufficient dimension reduction. *Ann. Statist.* **41**(1), 250–268.

prendergast, l. (2005) Influence functions for sliced inverse regression. *Scand. J. Stat.* **32**(3), 385–404.

prendergast, l. (2006) Detecting influential observations in sliced inverse regression analysis. *Aust. N. Z. J. Stat.* **48**(3), 285–304.

prendergast, l. (2007) Implications of influence function analysis for sliced inverse regession and sliced average variance estimation. *Biometrika* **94**(3), 585–601.

ripley, b.d. (1996) *Pattern Recognition and Neural Networks.*, Cambridge University Press, New York, NY.

rousseeuw, p.j. and leroy, a.m. (1987) *Robust Regression and Outlier Detection*, Wiley & Sons, New York, NY.

sheather, s.j. and mckean, j.w. (1997) A comparison pf procedures based on inverse regression. *IMS Lecture Notes – Monograph Series* **31**, 271–278.

sheather, s.j. and mckean, j.w. (2001) Discussion on: Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.* **43**(2), 185–190.

stewart, g.w. and sun, j. (1990) *Matrix Perturbation Theory*, 2nd ed. Academic Press, San Diego.

stromberg, a.j. and ruppert, d. (1992) Breakdown of nonlinear regression. *J. Amer. Statist. Assoc.* **87**, 991–997.

tyler, d.e. (2005) Discussion of “Breakdown and Groups” by P.L. Davies and U. Gather. *Ann. Statist.* **33**(3), 1009–1015.

welsh, a.h. (2001) Discussion on: Dimension reduction and visualization in discriminant analysis. *Aust. N. Z. J. Stat.* **43**(2), 177–179.

xia, y., tong, h., li, w.k. and zhu, l.-x. (2002) An adaptive estimation of dimension reduction space (with discussion). *J. R. Statist. Soc. B.* **64**, 363–410.

yin, x. and hilafu, h. (2014) Sequential sufficient dimension reduction for large p, small n problems. *J. R. Statist. Soc. B.* doi: 10.1111/rssb.12093.

Ulrike Genschel
Department of Statistics,
Iowa State University,
Ames, IA, USA
E-mail: ulrike@iastate.edu

Paper received: 24 March 2015.