Robustness Concepts for Sliced Inverse Regression

Dissertation

zur Erlangung des Grades

einer Doktorin der Naturwissenschaften

der Universität Dortmund

Dem Fachbereich Statistik

der Universität Dortmund

vorgelegt von

Ulrike Genschel

Dortmund, im April 2005

1. Gutachter: Prof. Dr. Ursula Gather

2. Gutachter: Prof. Dr. Claudia Becker

Tag der mündlichen Prüfung: 20. Juni 2005

Contents

Chapter 1

Introduction

Unquestionably, the need of dimension reduction procedures in statistical data analysis has never been more important than in today's time. Due to increased computing power and storage capacity of computers, the amount of data along with the number of surveyed variables in studies are often larger than ever before. Consider, for example, studies in the field of Bioinformatics or Epidemiology dealing with microarray data. When analyzing microarray data, we often face up to thousands of genes, representing regressor variables, but only very small sample sizes, as described by Tibshirani (2000) or Dettling and Bühlmann (2004) for example. The latter authors also refer to this as the "small sample size n, large predictor dimension p-phenomenon." But data sets consisting of fewer regressor variables can also pose severe problems in the data analysis step, which is the case in the often cited *Boston Housing* data set by Harrison and Rubinfeld (1978) that consists of 14 variables and $n = 506$ observations. The arising problem is known as the *curse of dimensionality* describing the following phenomena. As the dimension p of the regressor space increases, the space becomes sparser unless the amount of data, the number of observations n, grows exponentially. Consequently, nonparametric regression procedures used for fitting regression models will fail to estimate an underlying regression function g sufficiently well, as they crucially depend on data observations having neighboring observations nearby in the regression space.

As a remedy, often an appropriate lower dimensional *subspace* of the original regressor space is sufficient for fitting a regression function. These subspaces can be obtained by identifying $\mathcal{K} < p$ important linear combinations of the regressor variables.

SLICED INVERSE REGRESSION (SIR) by Li (1991) is a dimension reduction procedure that aims at estimating such a subspace, spanned by important linear combinations of regressor variables. Other well-known dimension reduction procedures estimating so-called dimension reduction subspaces of the regressor space include SIR II by Li (1991), SAVE (Cook and Weisberg (1991), Cook (2000)), PRINCIPAL HESSIAN DIRECTIONS (Li (1992)) or more recently MAVE by Xia et al. (2002). The main focus of this thesis though will only be on the dimension reduction procedure SIR.

Because ultimately the estimation of a regression curve or link function relies crucially on the correct identification of the linear combinations that span the dimension reduction subspace, robustness properties of a dimension reduction procedure become crucial to understand. That is, it is important to consider just how sensitive SIR and its subspace estimates are to data contamination. Unfortunately, there has been disagreement over the robustness properties of SIR.

Although Li (1991) pursues the argument that SIR is robust against outlying observations, other researchers including Hilker (1997), Gather et al. (2002) and Prendergast (2004) have demonstrated that the procedure is sensitive to certain types of data contamination which may influence the subspace estimate. Li (1991) justified his argument by pointing out that the influence of outlying observations within the response variable y is limited as y is incorporated in the SIR procedure only in a slicing step (see Chapter 3). Regarding the regressor variables x, he argues that x-values are typically fixed design points. However, as noted by Prendergast (2004), we are often faced with high dimensional data sets that are not the result of a designed experiment. Hence, we are obliged to be concerned about contamination of the data, in particular as such observations often remain undetected due to the high dimensionality of the data set [cf. Rousseeuw and Leroy (1987)]. Cook and Critchley (2000) reach the conclusion that outlying observations only result in additional linear combinations to the

true subspace. While this certainly may be true for some situations and particular types of contamination, we will show that outlying observations can indeed lead to the *loss* of important linear combinations of the subspace as well as the inclusion of *false* information (see Chapter 4). Bond (1999) also investigated the robustness of SIR focusing on the robustness of the procedure under violations of distributional assumptions. His study revealed that SIR is "fairly robust, not being affected by even fairly sizable perturbations when the response function was well behaved." We argue that these differing findings may arise because the *type of contamination* plays a significant role in the study of robustness of SIR. Single data points may not yield a bad estimate of a dimension reduction subspace simply if the points lie far away from the main bulk of the data. The direction in which outlying data points may play a more decisive and important role [cf. Hilker (1997), Gather et al. (2002)].

The focus of this thesis is placed on a detailed investigation of the robustness properties of SIR. In particular, we emphasize on the finite sample behavior of the SIR procedure under data contamination, considering various types of contamination (i.e., directions of contamination) which may produce a *worst case* subspace estimate. We wish to demonstrate that the data contamination scenarios that produce bad subspace estimates in SIR depend also on the covariance structure of the regressor variables as well as the knowledge of the dimension $\mathcal{K}$ of the final dimension reduction subspace.

Starting point is the dissertation of Hilker (1997), which provided a first thorough investigation of the robustness of SIR particulary focusing on a breakdown point definition. While this study of Hilker concentrated on the estimation of the first dimension reduction direction, we extend our study to all directions of the dimension reduction subspace. That is, we wish to consider the effect of data contamination on the *entire* dimension reduction *subspace* estimate as a whole, not single dimension reduction directions or vectors. As the dissertation of Prendergast (2004) offers first results on a definition of the influence function for SIR, we shall keep the main focus of our work on a possible definition of a finite sample breakdown point for SIR.

The thesis is organized as follows.

In **Chapter 2** we provide a review of necessary basic algebraic terminology allowing us to define distance measures between $\mathbb{R}^p$-vector subspaces. Our findings cumulate in an adequate metric on $\mathbb{R}^p$-subspaces that is based on the Frobenius norm. We will discuss its main properties and establish its suitability compared to other distance measures based on angles and so-called gap functions.

A review of the SIR procedure is given in **Chapter 3**. Because the procedure has been thoroughly presented in various sources in the literature, we will restrict our review only to details necessary for the investigation of the robustness properties.

The main theoretical contributions regarding the behavior of the SIR procedure under data contamination are presented in **Chapter 4**. We begin with a discussion of the meaning of the notion of *breakdown* with respect to SIR and dimension reduction procedures in general. This includes a review of previous approaches found in the literature, including those of Hilker (1997) and Becker (2001). We continue with a thorough analysis of the sensitivity of SIR, where a distinction of four different scenarios and data contamination schemes is necessary. We distinguish whether the dimension $\mathcal{K}$ of the dimension reduction subspaces to be estimated by the SIR procedure is known or unknown. A second decisive factor in the sensitivity analysis turns out to be the knowledge regarding the covariance matrix of the regressor variables. That is, we shall see that types of data contamination that cause SIR to yield an erroneous subspace estimate can change depending on whether the covariance of the regressors is known or not.

Chapter 5 contains the description of a simulation study used to numerically support our theoretical findings presented in the previous Chapter 4. The results of the simulation study concerning data contamination with SIR are presented and summarized.

The concept of BREAKDOWN & GROUPS established by Davies and Gather (2002, 2005a) offers by far the most insightful analysis regarding breakdown of a statistical functional. In **Chapter 6** we examine the extent to which the theory of BREAKDOWN & GROUPS can be applied to the case of dimension reduction functionals, in particular

to functionals of the SIR-type. We find that SIR-type functionals are difficult to place in the framework of BREAKDOWN & GROUPS, because a geometrically meaningful metric on subspaces (i.e., the parameters) is complicated to formulate under conditions required in Davies and Gather (2002, 2005a). As hinted by Davies and Gather (2004, 2005b), an alternative definition of breakdown can be applied to dimension reduction functionals which may statistically be more meaningful and accurate than a definition based on common breakdown point theory, since this requires no metric on subspaces to be specified.

Conclusions and recommendations for possible research extending the results found in this thesis are given in **Chapter 7**.

Some supporting technical arguments and results are provided in the **Appendix**.

Chapter 2

Preliminaries

This chapter focuses on Euclidian vector spaces $\mathbb{R}^p$ and developing metrics to measure the *distance* between subspaces of $\mathbb{R}^p$. In Section 2.1, a review is given of the basic properties of vector spaces and orthogonal projection matrices. Section 2.2 develops a metric for $\mathbb{R}^p$-subspaces based on the Frobenius matrix norm that has many desirable qualities for quantifying the distance between subspaces. For example, this subspace metric is often simple to compute and to geometrically interpret. We use this metric to formulate most of the finite sample breakdown results in Chapter 4. For completeness in Section 2.3, we discuss other general metrics which are possible for $\mathbb{R}^p$-subspaces, namely gap functions (Section 2.3.1) and angles (Section 2.3.2). We argue that the Frobenius norm-based subspace metric is more attractive than these other alternatives for subspace metrics. Section 2.4 illustrates the Frobenius norm-based metric with some examples. In Section 2.5, further properties of the Frobenius norm-based subspace metric are developed for later use, including the invariance of the metric (Section 2.5.1) and bounds on the maximal values of the metric (Section 2.5.2).

2.1 Vector Spaces

We begin by recalling some basic definitions involving Euclidean vector spaces $\mathcal{V} = \mathbb{R}^p, p \geq 1$ and subspaces as well as basis vectors which span them. Although these definitions concern basic algebraic elements, we feel that their review is worthwhile because SIR is concerned with estimating a vector subspace. The familiar reader may not hesitate to skip this section.

Definition 2.1 (Schott, 1995) *Let $\mathcal{V}$ denote a collection of $p \times 1$ vectors satisfying the following two properties*

P1. If $x_1 \in \mathcal{V}$ and $x_2 \in \mathcal{V}$, then $x_1 + x_2 \in \mathcal{V}$.

P2. If $x \in \mathcal{V}$ and α is a real-valued scalar, then $\alpha x \in \mathcal{V}$.

*Then $\mathcal{V}$ is called a **vector space** in p-dimensional space.*

Properties *P1* and *P2* are generally known as the closure properties for vector addition and scalar multiplication, respectively. In particular, the Euclidean vector space $\mathcal{V} = \mathbb{R}^p$ consists of all p-dimensional vectors with p real-valued components.

Definition 2.2 (Schott, 1995) *Let $\{x_1, \ldots, x_n\}$ be a set of $p \times 1$ vectors in a vector space $\mathcal{V}$. This set is called a **basis** of $\mathcal{V}$ if it spans the vector space $\mathcal{V}$ (i.e. any vector in $v \in \mathcal{V}$ can be expressed as a linear combination of $\{x_1, \ldots, x_n\}$ by $v = \sum_{i=1}^{n} \alpha_i x_i$ for real-valued scalars) and the vectors $x_1, \ldots, x_n$ are linearly independent (i.e. if $0 = \sum_{i=1}^{n} \alpha_i x_i$, then each $\alpha_i = 0$). Although this basis is not uniquely defined for a vector space, the number of vectors n in the basis is unique and is referred to as the **dimension** of the vector space $\mathcal{V}$, denoted by $\dim(\mathcal{V}) = n$.*

Definition 2.3 (Schott, 1995) *Let $\mathcal{V}$ be a vector space spanned by a basis $\{x_1, \ldots, x_n\}$. A space $\mathcal{S}$ is called a **subspace** of $\mathcal{V}$ ($\mathcal{S} \subset \mathcal{V}$) if it is spanned by a set $\{y_1, \ldots, y_{\mathcal{K}}\}, \mathcal{K} \leq n$, where any $y_i, i = 1, \ldots, \mathcal{K}$ is a linear combination of the $x_i's, i = 1, \ldots, n$ and $y_1, \ldots, y_{\mathcal{K}}$ are linearly independent. The number of linearly independent vectors $\mathcal{K}$ spanning $\mathcal{S}$ corresponds to the dimension of the subspace $\mathcal{S}$.*

Definition 2.4 *(Meyer, 2000)* *Let $\mathcal{S}_1$ and $\mathcal{S}_2$ be subspaces of a vector space $\mathcal{V}$. The **sum** of $\mathcal{S}_1$ and $\mathcal{S}_2$ is defined as the set of all possible sums of vectors from $\mathcal{S}_1$ with vectors from $\mathcal{S}_2$. That is*

$$\mathcal{S}_1 \oplus \mathcal{S}_2 = \{s_1 + s_2 \mid s_1 \in \mathcal{S}_1 \text{ and } s_2 \in \mathcal{S}_2\}.$$

Subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ generate a new subspace $\mathcal{S}_1 \oplus \mathcal{S}_2 = \widetilde{\mathcal{S}}$ and for two sets S_1 and S_2 of vectors spanning $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively. It follows that $S_1 \cup S_2$ spans $\mathcal{S}_1 \oplus \mathcal{S}_2 = \widetilde{\mathcal{S}}$.

In our studies of SIR and Euclidean vector subspaces, we will often refer to the orthogonal complement of a vector subspace. Two (nonzero) Euclidean vectors x and y are said to be orthogonal if the cosine of the angle between the vectors,

$$\cos \theta = \frac{x^\top y}{\|x\|\|y\|}, \tag{2.1}$$

equals zero. This leads to a general definition of orthogonality between Euclidean vectors (i.e. x and y are orthogonal if $x^\top y = 0$) and the notion of the orthogonal complement of a subspace, described in the following.

Definition 2.5 *(Schott, 1995)* *Let $\mathcal{S}$ be a Euclidean vector subspace of $\mathcal{V} = \mathbb{R}^p, p \geq 1$. The orthogonal complement of $\mathcal{S}$, denoted by $\mathcal{S}^\perp$, is the collection of all vectors in $\mathbb{R}^p$ that are **orthogonal** to every vector in $\mathcal{S}$; that is $\mathcal{S}^\perp = \{x \in \mathbb{R}^p : x^\top y = 0 \text{ for all } y \in \mathcal{S}\}$.*

In particular, assuming we have $\mathcal{V} = \mathbb{R}^p$, then if $\mathcal{S}$ is a vector subspace of $\mathbb{R}^p$ its orthogonal complement $\mathcal{S}^\perp$ is also a vector subspace of $\mathbb{R}^p$ and it holds that $\mathcal{V} = \mathbb{R}^p = \mathcal{S} \oplus \mathcal{S}^\perp$ (cf. Schott (1995), Theorem 2.15, p.56). It also holds that if a vector subspace $\mathcal{S}$ of $\mathbb{R}^p$ is of dimension $\mathcal{K}$, then the dimension of $\mathcal{S}^\perp$ is $p - \mathcal{K}$ (cf. Schott (1995), Theorem 2.16, p.56).

Our study of the robustness properties of SIR relies heavily on the notation of vector space projections and projection matrices involving Euclidean vector spaces $\mathcal{V} = \mathbb{R}^p$, $p \geq 1$. We review some necessary definitions in the following.

Definition 2.6 *(Fraleigh & Beauregard, 1990) Let $\mathcal{S}$ be a vector subspace of $\mathbb{R}^p$ with basis vectors $s_1, \ldots, s_{\mathcal{K}}$, $\mathcal{K} \leq p$, and let x be an arbitrary vector in $\mathbb{R}^p$. A vector $\widetilde{x} \in \mathbb{R}^p$ is said to be the **projection** of x onto $\mathcal{S}$ if and only if two properties hold:*

P1. $\widetilde{x} \in \mathcal{S}$,   i.e., $\widetilde{x} = \sum_{i=1}^{\mathcal{K}} a_i s_i$, for some $a_i \in \mathbb{R}$, $i = 1, \ldots, \mathcal{K}$.

P2. the vector $\widetilde{x} - x$ is orthogonal to $\mathcal{S}$,   i.e., $s^\top(\widetilde{x} - x) = 0$ for any $s \in \mathcal{S}$.

*A $p \times p$ matrix $P_{\mathcal{S}}$ is said to be the **orthogonal projection matrix** of the vector space $\mathcal{S}$ if $P_{\mathcal{S}} x = \widetilde{x}$ gives the projection of an arbitrary vector $x \in \mathbb{R}^p$ onto $\mathcal{S}$.*

Let $\mathcal{S} \subset \mathbb{R}^p$ be a vector subspace with basis vectors in $p \times \mathcal{K}$ matrix form $S = [s_1 \cdots s_{\mathcal{K}}]$, $\mathcal{K} \leq p$, whereby $\text{rank}(S) = \dim(\mathcal{S})$. To ease the exposition to follow, we will use a small abuse of notation and refer to either P_S (i.e., subscripted by the matrix S) or $P_{\mathcal{S}}$ (i.e., subscripted by the subspace $\mathcal{S}$ spanned by the columns of S) as the projection matrix for the subspace $\mathcal{S}$. This will cause no confusion because a subspace $\mathcal{S}$ is completely characterized both by its projection matrix $P_{\mathcal{S}} = P_S$ and by its basis. We will often make use of the following basic properties of projection matrices, which can be found in Schott (1995) or verified from Definition 2.6.

Properties of Projection matrices for a Euclidean Vector Space $\mathcal{S} \subset \mathbb{R}^p$:

P1. The matrix $P_{\mathcal{S}}$ is uniquely defined (although basis vectors S may not be) and may be written $P_{\mathcal{S}} = P_S = S(S^\top S)^{-1} S^\top$, where $\mathcal{K} = \text{rank}(S)$ for a matrix S with columns that are basis vectors spanning $\mathcal{S}$.

P2. For any $x \in \mathbb{R}^p$, the projection $P_{\mathcal{S}} x = \widetilde{x}$ of x onto $\mathcal{S}$ is unique.

P3. $P_{\mathcal{S}}$ is symmetric and also idempotent, i.e. $P_{\mathcal{S}}^\top = P_{\mathcal{S}}$ and $P_{\mathcal{S}}^2 = P_{\mathcal{S}}$.

P4. $\dim(\mathcal{S}) = \text{rank}(S) = \text{tr}(P_{\mathcal{S}})$.

P5. If $\widetilde{\mathcal{S}}$ is a further subspace of $\mathcal{S}$, namely $\widetilde{\mathcal{S}} \subset \mathcal{S} \subset \mathbb{R}^p$, then the projection matrix of the subspace $\mathcal{S} \cap \widetilde{\mathcal{S}}^\perp = \{x \in \mathbb{R}^p : x \in \mathcal{S}, \ \forall s \in \widetilde{\mathcal{S}}, \ x^\top s = 0\}$ equals $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$.

P6. If $x \in \mathbb{R}^p$ and $\mathcal{S} \subset \mathbb{R}^p$, then $\|x\|_2^2 = \|P_{\mathcal{S}} x\|_2^2 + \|(\mathrm{I}_p - P_{\mathcal{S}})x\|_2^2$, where I_p denotes the $p \times p$ identity matrix.

2.2 Defining distances between subspaces of $\mathbb{R}^p$

2.2.1 Metrics

Any attempt to formulate breakdown concepts for dimension reduction procedures, like SIR, will require us to develop a distance measure, or a metric, defined on subspaces of $\mathbb{R}^p$. We next recall the properties of a metric, defined on a general set of objects X.

Definition 2.7 *(**Mathieu, 1998**) Let X be a nonempty set. A nonnegative mapping $d : X \times X \to \mathbb{R}^p$ is called a **metric** on X, if the following properties hold*

P1. $\forall x, y \in X : d(x, y) = d(y, x)$.

P2. $\forall x, y, z \in X : d(x, y) \leq d(x, z) + d(z, y)$.

P3. $\forall x, y \in X : d(x, y) = 0 \Leftrightarrow x = y$.

*Under P1.-P3., the combination (X, d) is called a **metric space**.*

A set X coupled with a metric d determines a metric space (X, d). For every normed space $(X, \| \cdot \|)$ consisting of a set X and a norm $\| \cdot \|$, it is possible to define a metric space in a natural way by the distance measure $d(x, y) := \|x - y\|, x, y \in X$.

2.2.2 A metric induced by the Frobenius norm $\mathbb{F}$

It is well known that an examination of the robustness of a statistical procedure requires the definition of an appropriate metric on a parameter space (see Davies & Gather (2002, 2005a)). The main aim of a SIR-type method is to estimate a (dimension reduction) subspace of $\mathbb{R}^p$ spanned by certain p-dimensional basis vectors $\beta_1, \ldots, \beta_{\mathcal{K}}$. This subspace, $\text{span}(\beta_1, \ldots, \beta_{\mathcal{K}})$, may be considered as the parameter of interest in a parameter space consisting of all possible subsets of $\mathbb{R}^p$. Therefore, the

study of the robustness properties of SIR requires a suitable *metric* to assess distances between given $\mathbb{R}^p$-*subspaces*. Using an appropriately defined metric, we can quantify the amount of data contamination required to cause SIR to produce an erroneous and worst-case estimate of span($\beta_1, \ldots, \beta_{\mathcal{K}}$) (i.e., so-called breakdown). In the following, we define a useful metric on subspaces of $\mathbb{R}^p$ based on the concept of orthogonal projection matrices, which are common in the statistical literature for formulating ideas of subspaces and distances (e.g. Krzanowski (1979), Golub and Van Loan (1985), Crone and Crosby (1995) or Ferré (1998)).

Al previously noted, a vector subspace $\mathcal{S} \subset \mathbb{R}^p$ is completely characterized by its unique $p \times p$ projection matrix, $P_{\mathcal{S}}$. That is, there is an injective mapping, or one-to-one correspondence, from the set of $\mathbb{R}^p$-subspaces to the set of $p \times p$ real-valued matrices defined by $\mathcal{S} \mapsto P_{\mathcal{S}}$, with unique inverse $P_{\mathcal{S}} \mapsto \mathcal{S}$. This suggests that we can naturally judge the distance between two spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ by appropriately quantifying the distance between the corresponding matrices $P_{\mathcal{S}}$ and $P_{\widetilde{\mathcal{S}}}$. Such an approach allows us a geometrically attractive and mathematically tractable way to compare and measure distances between subspaces of $\mathbb{R}^p$, which may be applicable even to $\mathbb{R}^p$-subspaces of *different* dimensions. In fact, the comparison of subspaces of different dimensions can be especially useful and practical for a meaningful and logical concept of breakdown of a dimension reduction procedure. The idea is that projection matrices allow a comparison of the distance between $\mathbb{R}^p$-subspaces of *different* "dimensions" through matrices which are of the *same* size or "dimension" (namely $p \times p$). This can be seen easily in the following.

Suppose matrices $S = [s_1, \ldots, s_{\mathcal{K}}]$, $s_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{\mathcal{K}^*}]$, $\widetilde{s}_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}^*$ constitute an orthonormal basis for the (column) vector subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ of $\mathbb{R}^p$, respectively. Hence, the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are of dimensions $\mathcal{K}$ and $\mathcal{K}^*$ respectively, where $\mathcal{K}, \mathcal{K}^* \leq p$ and it may hold that $\mathcal{K} \neq \mathcal{K}^*$. Even for the $\mathcal{K} \neq \mathcal{K}^*$ case, the corresponding projection matrices $P_{\mathcal{S}} = S(S^\top S)^{-1}S^\top$ and $P_{\widetilde{\mathcal{S}}} = \widetilde{S}(\widetilde{S}^\top \widetilde{S})^{-1}\widetilde{S}^\top$ will naturally be $p \times p$ matrices. To induce a suitable *metric* between two vector spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ of $\mathbb{R}^p$, we will use the following two-step approach based on

$p \times p$ matrices:

1. compute the difference $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$

2. apply a matrix norm on $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$ to measure the *closeness* of $P_{\mathcal{S}}$ and $P_{\widetilde{\mathcal{S}}}$.

The idea is that we obtain a distance measure between vectors spaces, that has all the necessary properties of a metric, by using a metric on a special matrix $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$. For quantifying vector space distances in terms of projection matrix differences, we contend that a good choice of a matrix norm is the Frobenius norm defined below.

Definition 2.8 *(Stewart & Sun, 1990)* *Let $S \in \mathbb{R}^{m \times n}$ be an $m \times n$ matrix with components indexed by s_{ij}, $i = 1, \ldots, m$; $j = 1, \ldots, n$. The **Frobenius norm** of S is defined as*

$$\|S\|_{\mathbb{F}} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |s_{ij}|^2} = \sqrt{\mathrm{tr}(SS^{\top})}.$$

Stewart & Sun (1990, Chapter II, Section 2.1) describe basic properties of the norm $\| \cdot \|_{\mathbb{F}}$ and we state some useful properties of $\| \cdot \|_{\mathbb{F}}$ in the Appendix (see Lemma 7.4) for later use and reference.

We now give a precise definition of a metric $\mathbb{F}$ between two vector spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ based on the Frobenius norm.

Definition 2.9 *(Frobenius norm-based metric definition on subspaces)* *Let matrices $S = [s_1, \ldots, s_{\mathcal{K}}]$, $s_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{\mathcal{K}^*}]$, $\widetilde{s}_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}^*$ represent an orthonormal basis for the (column) vector subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively. The distance between $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ using the Frobenius norm $\| \cdot \|_{\mathbb{F}}$ is then defined by*

$$
\begin{aligned}
\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \|P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}\|_{\mathbb{F}} &= \left[\mathrm{tr}\{(P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}})(P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}})^{\top}\} \right]^{\frac{1}{2}} \\
&= [\mathrm{tr}(P_{\mathcal{S}}^2 - P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}} - P_{\widetilde{\mathcal{S}}}P_{\mathcal{S}} + P_{\widetilde{\mathcal{S}}}^2)]^{\frac{1}{2}} \\
&= [\mathrm{tr}(P_{\mathcal{S}}) + \mathrm{tr}(P_{\widetilde{\mathcal{S}}}) - 2\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}})]^{\frac{1}{2}} \\
&= [\mathrm{rank}(S) + \mathrm{rank}(\widetilde{S}) - 2\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}})]^{\frac{1}{2}} \qquad (2.2) \\
&= [\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - 2\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}})]^{\frac{1}{2}}.
\end{aligned}
$$

Because all s_i, $i = 1, \ldots, \mathcal{K}$ and $\widetilde{s}_i$, $i = 1, \ldots, \mathcal{K}^*$ are orthonormal, the unique projection matrices onto the vector spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are given by $P_S = S(S^\top S)^{-1} S^\top = SS^\top$ and $P_{\widetilde{S}} = \widetilde{S}(\widetilde{S}^\top \widetilde{S})^{-1} \widetilde{S}^\top = \widetilde{S}\widetilde{S}^\top$, respectively.

In the above definition of $\mathbb{F}$ we use the property that the trace of a projection matrix $P_{\mathcal{S}}$ equals the rank or dimension of the subspace $\mathcal{S}$. The metric $\mathbb{F}$ also inherently incorporates the term $\mathrm{tr}(P_S P_{\widetilde{S}})$, which may be viewed as a measure of the closeness of two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, a nonnegative numerical measure of what both subspaces have in common. Lemma 7.3 shows that $\mathrm{tr}(P_S P_{\widetilde{S}}) \geq 0$ and this trace equals zero only if the $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal, sharing only the zero vector $0_p \in \mathbb{R}^p$ in common. As the size or dimension of the intersection $\mathcal{S} \cap \widetilde{\mathcal{S}}$ increases, so will the corresponding trace $\mathrm{tr}(P_S P_{\widetilde{S}})$ term. Krzanowski (1979) used this same trace quantity for assessing the closeness when studying the subset of principal components in multivariate analysis. However, the trace $\mathrm{tr}(P_S P_{\widetilde{S}})$ does not constitute a metric by failing to satisfy a triangle inequality. The form in (2.2) however ensures that $\mathbb{F}$ immediately defines a metric on vector spaces; see Section 2.2.1. This fact is also illustrated in the following properties.

Metric Properties of $\mathbb{F}$: For vector spaces $\mathcal{S}, \widetilde{\mathcal{S}}, \mathcal{T} \subset \mathbb{R}^p$, it holds that

P1. $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) \geq 0$ (nonnegativity).

P2. $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = 0 \Longleftrightarrow P_{\mathcal{S}} = P_{\widetilde{\mathcal{S}}} \Longleftrightarrow \mathcal{S} = \widetilde{\mathcal{S}}$ (identifiability).

P3. $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \mathbb{F}(\widetilde{\mathcal{S}}, \mathcal{S})$ (symmetry).

P4. $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \mathbb{F}(\mathcal{T}, \mathcal{S}) + \mathbb{F}(\mathcal{T}, \widetilde{\mathcal{S}})$ (triangle inequality).

The second property above also holds true up to an orthogonal transformation of the vector space $\mathcal{S}$ or equivalently an orthogonal transformation of the basis vectors S (in matrix form) of $\mathcal{S}$; the invariance of the metric $\mathbb{F}$ to orthogonal transformations of subspaces follows from the invariance of the Frobenius matrix norm $\| \cdot \|_{\mathbb{F}}$ to orthogonal transformations of either the rows or columns of a matrix (see Lemma 7.4(d) in

the Appendix for this matrix norm property). We shall further detail the invariance property of $\mathbb{F}$ in Section 2.5.1.

Using the Frobenius matrix norm, Paige (1984) proposed a distance measure on subspaces $\mathcal{S}, \widetilde{\mathcal{S}} \subset \mathbb{R}^p$ of the *same* dimension $\dim(\mathcal{S}) = \dim(\widetilde{\mathcal{S}}) = \mathcal{K}$ given by a metric:

$$\min_{\mathcal{K} \times \mathcal{K} \text{ orthonormal } Q} \|S - \widetilde{S}Q\|_{\mathbb{F}} = \|(P_S - P_{\widetilde{S}})S\|_{\mathbb{F}},$$

where the columns of the $p \times \mathcal{K}$ matrices $S, \widetilde{S}$ must constitute orthonormal bases of $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively, and the minimum above is achieved at the $\mathcal{K} \times \mathcal{K}$ orthogonal matrix $Q = \widetilde{S}^\top S$ yielding the equality above. We remark that Stewart & Sun (1990, Chapter II, Section 4) consider this metric along with others on $\mathbb{R}^p$-subspaces of the same dimension which we will describe in Section 2.3.

For measuring distances between subspaces in hypothesis tests with principal-component regression, Crone & Crosby (1995) also employed a metric similar to $\mathbb{F}$. Given the principal components associated with the $\mathcal{K}$ largest eigenvalues of two different $p \times p$ sample covariance matrices and writing those $\mathcal{K}$ principal components as the columns of two $p \times \mathcal{K}$ matrices A and B of rank $\mathcal{K}$, these authors use $\left(\mathcal{K} - \text{tr}[A(A^\top A)^{-1}A^\top B(B^\top B)^{-1}B^\top]\right)^{1/2}$ as the distance between the two $\mathcal{K}$-dimensional subspaces of $\mathbb{R}^p$ given by the column spaces of matrices A and B. Applying (2.2) to the column space of matrices A and B gives the above distance measure of Crone & Crosby (1995), scaled by a factor of $\sqrt{2}$:

$$\left(\mathcal{K} - \text{tr}[A(A^\top A)^{-1}A^\top B(B^\top B)^{-1}B^\top]\right)^{\frac{1}{2}} = \frac{\mathbb{F}(\text{col}(A), \text{col}(B))}{\sqrt{2}}.$$

Hence, (2.2) may be viewed as a generalization of the special subspace metric of Crone & Crosby (1995) which allows us to accommodate general subspaces of $\mathbb{R}^p$ that may not have the same fixed dimension.

2.3 General distance measures

To further frame the choice of the subspace metric defined in (2.2), we shall briefly explain some other general distance measures between vector spaces which have been introduced in the literature. (We refer the interested reader to Stewart & Sun (1990) and Meyer (2000) for more details.)

There are two possible approaches for developing subspace metrics: either through gap functions between subspaces or through angles between subspaces. We will find that our subspace metric $\mathbb{F}(\cdot, \cdot)$ from (2.2) is related to so-called gap functions in Section 2.3.1 and in fact all subspace metrics based on gap functions are equivalent (i.e. they lead to the same topology on the set of all possible $\mathbb{R}^p$-subspaces). In Section 2.3.2, we argue that $\mathbb{F}(\cdot, \cdot)$ is more appropriate and tractable than angles to measure distances between subspaces.

2.3.1 Gap functions

The first approach to define an appropriate distance measure between two vector subspaces as in our case is to begin with the definition of the distance between a point and a subspace.

Definition 2.10 *(Stewart & Sun, 1990)* *Let $\mathcal{S}$ denote a subspace of $\mathbb{R}^p$ and b a point in $\mathbb{R}^p$. By ν we denote an arbitrary vector norm on $\mathbb{R}^p$. The ν-**distance** between $\mathcal{S}$ and b with respect to the norm ν is then defined as*

$$\operatorname{dist}_\nu(b, \mathcal{S}) = \min_{s \in \mathcal{S}} \nu(b - s). \tag{2.3}$$

Considering the special case of the Euclidian vector norm ($\nu(\cdot) = \|\cdot\|_2$) on $\mathbb{R}^p$, the above distance (2.3) can be expressed as

$$\operatorname{dist}_2(b, \mathcal{S}) = \min_{s \in \mathcal{S}} \|b - s\|_2 = \|b - P_{\mathcal{S}} b\|_2 = \|(\mathrm{I} - P_{\mathcal{S}}) b\|_2,$$

which corresponds to the *orthogonal distance* between b and the subspace $\mathcal{S}$ or equivalently the distance between b and its projection onto $\mathcal{S}$ [cf. Meyer (2000)]. We can then extend the orthogonal distance (2.3) between a point b and a subspace $\mathcal{S}$ to formulate the distance between two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ by the following computations

$$
\begin{aligned}
\text{dist}_2(\mathcal{S}, \widetilde{\mathcal{S}}) &= \max_{\substack{s \in \mathcal{S} \\ \|s\|_2=1}} \text{dist}_2(s, \widetilde{\mathcal{S}}) \\
&= \max_{\substack{s \in \mathcal{S} \\ \|s\|_2=1}} \|(I - P_{\widetilde{\mathcal{S}}})s\|_2 \\
&= \max_{\substack{s \in \mathcal{S} \\ \|s\|_2 \leq 1}} \|(I - P_{\widetilde{\mathcal{S}}})s\|_2 \qquad (2.4) \\
&= \max_{\substack{\|x\|_2=1 \\ x \in \mathbb{R}^p}} \|(I - P_{\widetilde{\mathcal{S}}})P_{\mathcal{S}}x\|_2. \qquad (2.5)
\end{aligned}
$$

Equation (2.4) can be justified because for a nonnegative function $f : \mathcal{V} \to \mathbb{R}$ defined on a vector space $\mathcal{V}$ such that $f(\alpha x) = \alpha f(x)$ for some scalar $\alpha \geq 0$

$$
\max_{\|x\|_2=1} f(x) = \max_{\|x\|_2 \leq 1} f(x)
$$

holds; (2.5) follows because $\{s \in \mathcal{S} : \|s\|_2 \leq 1\} = \{P_{\mathcal{S}}x : x \in \mathbb{R}^p, \|x\|_2 = 1\}$.

It should be noted that the orthogonal distance $\text{dist}_2(\mathcal{S}, \widetilde{\mathcal{S}})$ corresponds actually to a directed distance from $\mathcal{S}$ to the subspace $\widetilde{\mathcal{S}}$, which may not necessarily equal the orthogonal distance $\text{dist}_2(\widetilde{\mathcal{S}}, \mathcal{S})$ from $\widetilde{\mathcal{S}}$ to $\mathcal{S}$ (i.e. reversing the order $\widetilde{\mathcal{S}}, \mathcal{S}$). Hence, using the orthogonal distance $\text{dist}_2(\cdot, \cdot)$ from (2.5) to quantify the maximal degree of separation between two subspaces requires both directed distance values $\text{dist}_2(\mathcal{S}, \widetilde{\mathcal{S}})$ and $\text{dist}_2(\widetilde{\mathcal{S}}, \mathcal{S})$ to be incorporated in a more complete distance measure, the so-called GAP [cf. Meyer (2000), p. 453]:

$$
\text{gap}_2(\mathcal{S}, \widetilde{\mathcal{S}}) = \max \left\{ \text{dist}_2(\mathcal{S}, \widetilde{\mathcal{S}}), \text{dist}_2(\widetilde{\mathcal{S}}, \mathcal{S}) \right\}. \qquad (2.6)
$$

Note that the GAP measure (2.6) above grew out of orthogonal distance considerations (2.4) between a point and a subspace based on the Euclidean vector norm ($\nu(\cdot) = \|\cdot\|_2$) on $\mathbb{R}^p$. A more general definition for distances between subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$, based on the GAP concept, can be formulated by beginning with an arbitrary vector norm ν on $\mathbb{R}^p$.

Definition 2.11 *(Stewart & Sun, 1990)* *Let $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ denote two subspaces of $\mathbb{R}^p$. The GAP between $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ with respect to a vector norm ν on $\mathbb{R}^p$ is defined as*

$$\mathrm{gap}_\nu(\mathcal{S}, \widetilde{\mathcal{S}}) = \max \left\{ \max_{\substack{s \in \mathcal{S} \\ \nu(s)=1}} \mathrm{dist}_\nu(s, \widetilde{\mathcal{S}}), \max_{\substack{\widetilde{s} \in \widetilde{\mathcal{S}} \\ \nu(\widetilde{s})=1}} \mathrm{dist}_\nu(\widetilde{s}, \mathcal{S}) \right\} \tag{2.7}$$

with $\mathrm{dist}_\nu(\cdot, \cdot)$ defined as in (2.3).

We recall that the subspace metric $\mathbb{F}$ from Definition 2.9 involved applying the Frobenius matrix norm $\| \cdot \|_{\mathbb{F}}$ to the difference of the projection matrices $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$. For defining alternative metrics between $\mathbb{R}^p$- subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, we may note an interesting connection between the GAP function based on the Euclidean vector norm (2.6) and another matrix norm applied to the difference in projection matrices $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$ (but not the Frobenius matrix norm). The spectral norm or matrix 2-norm defined on a $q \times p$ real-valued matrix A is given as

$$\|A\|_2 = \max_{x \in \mathbb{R}^p, \|x\|_2 = 1} \|Ax\|_2 = \sqrt{\text{largest eigenvalue of } A^\top A}. \tag{2.8}$$

It can be shown that the GAP based on the Euclidean vector norm (2.6) can be reformulated in terms of the matrix 2-norm (2.8) applied to the difference of projection matrices for two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ of $\mathbb{R}^p$:

$$\mathrm{gap}_2(\mathcal{S}, \widetilde{\mathcal{S}}) = \|P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}\|_2. \tag{2.9}$$

In considering numerical methods for matrix computations, Golub and Van Loan (1985) apply this same distance measure (2.9) on subspaces $\mathcal{S}, \widetilde{\mathcal{S}} \subset \mathbb{R}^p$. The metric (2.9) on subspaces resembles the metric $\mathbb{F}$ from Definition 2.9 by applying the *spectral norm* (2.8), rather than the Frobenius norm, on the difference of projection matrices $P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}$.

We remark that the subspace metrics in (2.2), Definition 2.9 and (2.9) are special cases of a more general framework given in Stewart and Sun (1990, Chapter II, Theorem 4.7) for defining subspace metrics. Namely, with *any* matrix norm ν_m for $p \times p$ matrices, the function

$$\rho_{\nu_m}(\mathcal{S}, \widetilde{\mathcal{S}}) \equiv \nu_m(P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}) \tag{2.10}$$

yields a metric ρ_{ν_m} on subspaces of $\mathbb{R}^p$ [cf. Stewart & Sun (1990), Chapter II, Theorem 4.7; though these authors consider only $\mathbb{R}^p$-subspaces of the same dimension]. Possibilities for subspace metrics $\rho_{\nu_m}(\mathcal{S}, \widetilde{\mathcal{S}})$ in (2.10) could also involve a norm $\nu_m(A)$ on $p \times p$ matrices A defined by

$$\nu_m(A) = \|A\|_q = \max_{x \in \mathbb{R}^p, \|x\|_q = 1} \|Ax\|_q$$

based on a Hölder vector norm $\|\cdot\|_q$ for $q \in \mathbb{N} \cup \{\infty\}$; note again that (2.8) is equivalent to using $\|\cdot\|_2$ in (2.10). However, all subspace metrics defined by gap functions (2.10), including $\mathbb{F}$, are in some sense equivalent because of the equivalence of matrix norms [cf. Stewart and Sun (1990), Chapter II, Section 4]. In particular, if ν_m and $\widetilde{\nu}_m$ are $p \times p$ matrix norms defined by $\|A\|_{\mathbb{F}}$ or $\|A\|_q$ for any $q \in \mathbb{N} \cup \{\infty\}$, then there exist positive constants $\alpha, \beta > 0$ such that

$$\frac{\alpha}{\beta} \cdot \rho_{\nu_m}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \rho_{\widetilde{\nu}_m}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \frac{\beta}{\alpha} \cdot \rho_{\nu_m}(\mathcal{S}, \widetilde{\mathcal{S}})$$

for all subspaces $\mathcal{S}, \widetilde{\mathcal{S}} \subset \mathbb{R}^p$. This follows because $\alpha\nu_m(A) \leq \widetilde{\nu}_m(A) \leq \beta\nu_m(A)$ holds for some constants $\alpha, \beta > 0$ for choices of matrix norms $\nu_m, \widetilde{\nu}_m$ from $\|\cdot\|_{\mathbb{F}}$ or $\|\cdot\|_q$, $q \in \mathbb{N} \cup \{\infty\}$. We refer to Lemma 7.4 for an example involving the matrix norms $\|\cdot\|_{\mathbb{F}}$ and $\|\cdot\|_2$. Hence, subspace metrics ρ_{ν_m} based on a Hölder matrix norm $\nu_m = \|\cdot\|_q$, $q \in \mathbb{N} \cup \{\infty\}$, will generate the same topology as $\mathbb{F}$ in (2.2) on subspaces of $\mathbb{R}^p$.

2.3.2 Angles

A common and perhaps more intuitive distance measure to gauge the separation between subspaces are angles. The major drawback from employing angles as appropriate measures, however, is that they often do not constitute a metric between subspaces. Furthermore, angles as distance measures for vector subspaces of $\mathbb{R}^p$ in higher dimensions p are not as straightforward or appropriate as in lower dimensional spaces $\mathbb{R}^2$ or $\mathbb{R}^3$, and certainly can be more difficult to visualize.

There are often several choices of angles between subspaces that are available to quantify the distance between subspaces but the suitability of a subspace angle type

may depend on whether the considered subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are of the same dimension or whether the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are complementary (i.e., $\mathcal{S} = \widetilde{\mathcal{S}}^{\perp}$, $\mathbb{R}^p = \mathcal{S} \oplus \widetilde{\mathcal{S}}$ for $\mathcal{S}, \widetilde{\mathcal{S}} \subset \mathbb{R}^p$). We shall also see that the subspace metrics we have established so far, like (2.2) and (2.9), incorporate angles automatically in their computation and thus actually incorporate more *angular* information on subspaces than one might think at first glance. These types of considerations are the topic of discussion in this section.

For $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, a first type of angle to measure the separation of the subspaces is the so-called MINIMAL ANGLE. The minimal angle $\theta_{\min}$ between subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ is defined as the value $0 \leq \theta_{\min} \leq \pi/2$ which satisfies

$$\cos \theta_{\min} = \max_{\substack{s \in \mathcal{S}, \widetilde{s} \in \widetilde{\mathcal{S}} \\ \|s\|_2 = \|\widetilde{s}\|_2 = 1}} s^{\top} \widetilde{s}. \tag{2.11}$$

If $P_{\mathcal{S}}$ and $P_{\widetilde{\mathcal{S}}}$ are the orthogonal projection matrices onto $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively, then the minimal angle can be computed as $\cos \theta_{\min} = \|P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}}\|_2$, applying the matrix 2-norm from (2.8). From the definition of the minimal angle, it is clear that two $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ will satisfy $\cos \theta_{\min} = 0$ if the two spaces are complementary. However, it always holds that $\cos \theta_{\min} = 1$ (the maximal value) whenever two $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ have a nontrivial intersection $\mathcal{S} \cap \widetilde{\mathcal{S}} \neq \{0_p\}$ [cf. Meyer (2000), p. 452]. It is then less obvious how to interpret $\cos \theta_{\min}$ as a distance measure if the two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ are non-complementary whenever $p > 2$. Hence, the application of this type of angle can become rather limited.

While the $\cos \theta_{\min}$ in (2.11) is defined as the minimal angle, $\sin \theta_{\max}$ describes the so-called MAXIMAL ANGLE $\theta_{\max}$ between $\mathbb{R}^p$ subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, which unlike the minimal angle fulfills all properties of a metric. It turns out that the sine function of the maximal angle is equal to the GAP function defined in (2.6), which takes on values between 0 and 1; that is, the maximal angle between two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ corresponds to the number $0 \leq \theta_{\max} \leq \pi/2$ satisfying

$$\sin \theta_{\max} = \mathrm{gap}_2(\mathcal{S}, \widetilde{\mathcal{S}}) = \|P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}\|_2. \tag{2.12}$$

While the minimal angle can measure the degree of closeness between two nontrivial complementary subspaces, the maximal angle is most suitable for subspaces of *equal*

dimension. This is because of the following properties of $\sin\theta_{\max}$, given by Meyer (2000) [p. 454]:

$$\|P_{\mathcal{S}} - P_{\widetilde{\mathcal{S}}}\|_2 = \sin\theta_{\max} = \begin{cases} 1 & \text{if } \dim(\mathcal{S}) \neq \dim(\widetilde{\mathcal{S}}), \\ 1 & \text{if } \dim(\mathcal{S}) = \dim(\widetilde{\mathcal{S}}), \mathcal{S} \cap \widetilde{\mathcal{S}}^{\perp} \neq 0_p, \\ c < 1 & \text{if } \dim(\mathcal{S}) = \dim(\widetilde{\mathcal{S}}), \mathcal{S} \cap \widetilde{\mathcal{S}}^{\perp} = 0_p. \end{cases}$$

Hence, (2.9) and (2.12) are equal to 1 (the maximal value) whenever two $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ have unequal dimension and this metric also equals 1 whenever $\mathcal{S}$ and $\widetilde{\mathcal{S}}^{\perp}$ share a non-zero vector in common.

We find that neither minimal nor maximal angles are appropriate as a measure of the amount of separation between two general subspaces. For example, if $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are $\mathbb{R}^p$-subspaces of unequal dimension that have a nontrivial intersection, then the minimal angle $\theta_{\min} = 0$ and the maximal angle $\theta_{\max} = \pi/2$, but neither number may convey the desired information of the closeness of the two subspaces.

To formulate angle values as a comprise between the extremes $\theta_{\min}$ and $\theta_{\max}$, the so-called PRINCIPAL or CANONICAL ANGLES have been derived in the literature to assess the distance between two subspaces. Canonical angles were also used for example by Hilker (1997) to measure the distance between contaminated subspace estimates in the study of SIR. For nonzero subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ with $m = \min\{\dim\mathcal{S}, \dim\widetilde{\mathcal{S}}\}$, the PRINCIPAL ANGLES between $\mathcal{S} = \mathcal{S}_1$ and $\widetilde{\mathcal{S}} = \widetilde{\mathcal{S}}_1$ are recursively defined to be the numbers $0 \leq \theta_i \leq \pi/2$ such that

$$\cos\theta_i = \max_{\substack{s \in \mathcal{S}_i, \widetilde{s} \in \widetilde{\mathcal{S}}_i \\ \|s\|_2 = \|\widetilde{s}\|_2 = 1}} s^{\top}\widetilde{s} = s_i^{\top}\widetilde{s}_i, \quad i = 1, \ldots, m,$$

where $\|s_i\|_2 = \|\widetilde{s}_i\|_2 = 1$, $\mathcal{S}_i = s_{i-1}^{\perp} \cap \mathcal{S}_{i-1}$, $\widetilde{\mathcal{S}}_i = \widetilde{s}_{i-1}^{\perp} \cap \widetilde{\mathcal{S}}_{i-1}$ with $\mathcal{S}_1 = \mathcal{S}, \widetilde{\mathcal{S}}_1 = \widetilde{\mathcal{S}}$. It then holds that $\theta_{\min} = \theta_1 \leq \cdots \leq \theta_m \leq \theta_{\max}$ with $\theta_m = \theta_{\max}$ in the case of $\dim\mathcal{S} = \dim\widetilde{\mathcal{S}}$ [cf. Meyer (2000), p. 456]. Hence, the principal angles may be viewed as intermediate angles between $\theta_{\min}$ and $\theta_{\max}$. The limitation of the principal angles as a measure of closeness between subspaces in the context of this paper becomes obvious because principal angles do not constitute a metric. Although they allow for a comparison of

subspaces of different dimensions, the number of principal angles that can be computed is at most $m = \min\{\dim \mathcal{S}, \dim \widetilde{\mathcal{S}}\}$, which is limited by the size or dimension of the *smallest* subspace in the comparison of two subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$.

To facilitate mathematical computation and geometrical interpretation, we prefer to use the metric $\mathbb{F}$ over (2.9) or other gap functions (2.10). This choice of a distance measure $\mathbb{F}$ also has desirable invariance properties with respect to orthogonal transformations; see Section 2.5.1. By using $\mathbb{F}$, we also avoid the complicated geometrical framework given by James (1954), who developed an approach for describing distributions of k-dimensional subspaces of $\mathbb{R}^p$ based on manifolds and differential geometry.

An appealing feature of the metric $\mathbb{F}$ is that the metric, in some sense, agrees with our geometrical intuition of the *closeness* between subspaces, while still constituting a *metric*. We next illustrate this quality with some numerical examples.

2.4   Numerical examples of $\mathbb{F}$

We begin this section with a numerical example. For the purpose of illustration we consider a simple example.

Example 2.1 *Let $\mathcal{V} = \mathbb{R}^8$ denote a vector space spanned by unit vectors e_i, $i = 1, \ldots, 8$ with e_i containing the entry 1 in the ith position and 0 everywhere else. Hence, $\|e_i\| = 1$ and $e_i^\top e_j = 0$ for $i \neq j$. Furthermore, we denote by $\mathcal{S}$ a subspace of $\mathbb{R}^8$, which for simplicity is spanned by five of the basis vectors, namely $e_1, \ldots, e_5$ ($\mathcal{K} = 5$). Suppose that a second subspace $\widetilde{\mathcal{S}}$ is spanned by some sub-collection of the vectors $e_1, \ldots, e_8$. (With regard to our application in SIR (see Chapter 3), $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ could represent the true and the estimated dimension reduction spaces, respectively.) We indicate by OL the number of OVERLAPPING directions between $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ or the numbers of vectors among $\{e_1, \ldots, e_8\}$ that are common to both $\mathcal{S}$ and $\widetilde{\mathcal{S}}$. We use NOL to denote the number of NONOVERLAPPING directions between $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ or the number of vectors*

$\dim(\mathcal{S})$	$\dim(\widetilde{\mathcal{S}})$	OL directions	NOL directions	$\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}})$
5	8	5	3	$\sqrt{3}$
5	7	5	2	$\sqrt{2}$
5	7	4	3	$\sqrt{4}$
5	6	5	1	$\sqrt{1}$
5	6	4	2	$\sqrt{3}$
5	6	3	3	$\sqrt{5}$
5	5	5	0	$\sqrt{0}$
5	5	4	1	$\sqrt{2}$
5	5	3	2	$\sqrt{4}$
5	5	2	3	$\sqrt{6}$
5	4	4	0	$\sqrt{1}$
5	4	3	1	$\sqrt{3}$
5	4	2	2	$\sqrt{5}$
5	4	1	3	$\sqrt{7}$
5	3	3	0	$\sqrt{2}$
5	3	2	1	$\sqrt{4}$
5	3	1	2	$\sqrt{6}$
5	3	0	3	$\sqrt{8}$
5	2	2	0	$\sqrt{3}$
5	2	1	1	$\sqrt{5}$
5	2	0	2	$\sqrt{7}$
5	1	1	0	$\sqrt{4}$
5	1	0	1	$\sqrt{6}$

TABLE 1: Computations of Frobenius norm based metric between subspaces of $\mathbb{R}^8$

among $\{e_1, \ldots, e_8\}$ *that are in the span of* $\widetilde{\mathcal{S}}$ *but not in the span of* $\mathcal{S}$. *In the first two columns of Table 1 the dimensions for* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$ *are listed. Applying the metric introduced in Definition 2.9, the distances in Table 1 between* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$ *are obtained. Obviously, the metric becomes only zero, when* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$ *are spanned by the same set of vectors, while it takes on its maximal distance* $\sqrt{8}$, *when* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}}$ *are completely orthogonal to each other.*

Example 2.2 *(subspaces spanned by lines in* $\mathbb{R}^p$*, i.e., of dimension 1)*

It is helpful to consider some simple cases to illustrate that the vector space metric $\mathbb{F}$ *agrees with geometric intuition in a sense. For example, suppose two vector spaces* $\mathcal{S}$ *and* $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ *correspond to two lines spanned by normalized vectors* s_1 *and*

$\widetilde{s}_1$, respectively (i.e., $\|s_1\| = \|\widetilde{s}_1\| = 1$). In this special case, the cosine of the angle $\theta = \cos^{-1}(s_1^\top \widetilde{s}_1)$ between the vectors s_1 and $\widetilde{s}_1$, namely $s_1^\top \widetilde{s}_1$, embodies a common and intuitively appealing measure of the closeness of the lines $\mathcal{S} = \mathrm{span}(s_1)$ and $\widetilde{\mathcal{S}} = \mathrm{span}(\widetilde{s}_1)$, corresponding to the minimal angle defined in (2.11). Applying the metric $\mathbb{F}$ to these 1-dimensional spaces, we find here that

$$\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2} \cdot \left[1 - \cos^2(\theta)\right]^{1/2} = \sqrt{2} \cdot [1 - (s_1^\top \widetilde{s}_1)^2]^{1/2} = \sqrt{2} \cdot \sin\theta.$$

That is, the metric $\mathbb{F}$ incorporates the same natural measure of distance between two lines in terms of the cosine of the canonical angle $\cos(\theta) = s_1^\top \widetilde{s}_1$ formed by the lines. When $\cos(\theta) = 1$, we have that the lines $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are the same and the metric $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = 0$ reflects this aspect as well by assuming a value of zero; at the other extreme, when the lines are orthogonal $s_1^\top \widetilde{s}_1 = 0$, the spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are further apart and the metric $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2}$ reflects this difference in distance as well.

In this case with two subspaces spanned by normalized vectors s_1 and $\widetilde{s}_1 \in \mathbb{R}^p$, it is interesting to contrast the computation of $\mathbb{F}$ with the evaluation of the matrix-2 norm based metric using its definition in (2.9). If $p = 2$, we can explicitly determine (2.9) as the square root of the largest root of the characteristic polynomial $\det(A - \lambda \mathrm{I}_p)$, with $A = (s_1 s_1^\top - \widetilde{s}_1 \widetilde{s}_1^\top)^2$ and $p \times p$ identity matrix I_p, given by

$$\lambda^2 - \lambda \mathrm{tr}(A) + \det(A) = 0 \quad \text{(if } p = 2\text{)}.$$

In the case of $p = 2$, it can be verified through direct computation that $\det(A) = (\mathrm{tr}(A)/2)^2$, so that the above characteristic polynomial is

$$\begin{aligned} 0 &= \lambda^2 - \lambda \mathrm{tr}(A) + \det(A) \\ &= (\lambda - \frac{\mathrm{tr}(A)}{2})^2, \end{aligned}$$

which has a single root at

$$\lambda = \frac{\mathrm{tr}(A)}{2} \geq 0.$$

The square root of this λ- value is the value of the metric (2.9):

$$
\begin{aligned}
\sin \theta_{\max} &= \operatorname{gap}_2(\mathcal{S}, \widetilde{\mathcal{S}}) \\
&= \sqrt{\frac{\operatorname{tr}(A)}{2}} \\
&= \sqrt{1 - \cos^2(\theta)} \\
&= \frac{\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}})}{\sqrt{2}}.
\end{aligned}
$$

That is, we find that the distance between $\mathcal{S} = \operatorname{span}(s_1)$ and $\widetilde{\mathcal{S}} = \operatorname{span}(\widetilde{s}_1)$ by the spectral metric (2.9) equals $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}})/\sqrt{2}$ and $\cos^{-1}(s_1^\top \widetilde{s}_1) = \theta = \theta_{\max}$. Even for this simple case, we see the computation of the distance (2.9) using the definition is much more involved and here embodies the same information as $\mathbb{F}$ in terms of the canonical angle $\cos^{-1}(s_1^\top \widetilde{s}_1)$ between $\mathcal{S}$ and $\widetilde{\mathcal{S}}$.

2.5 Further properties of the metric $\mathbb{F}$

2.5.1 Invariance

A first desirable property of the subspace metric $\mathbb{F}$ is its invariance to orthogonal transformations of a subspace $\mathcal{S} \subset \mathbb{R}^p$. Because a dimension reduction procedure like SIR attempts to estimate the dimension reduction space $\operatorname{span}(\beta_1, \ldots, \beta_{\mathcal{K}}) \subset \mathbb{R}^p$ only up to an orthogonal transformation of the vectors β_i, this invariance of the subspace metric $\mathbb{F}$ is particularly important. To demonstrate invariance for transformation of the columns (or rows) of a matrix, define matrices $S = [s_1, \ldots, s_{\mathcal{K}}]$, $s_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{\mathcal{K}^*}]$, $\widetilde{s}_i \in \mathbb{R}^p$ for $i = 1, \ldots, \mathcal{K}^*$ as orthonormal bases for the (column) subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively, and let $Q_{\mathcal{K}}$ and $Q_{\mathcal{K}^*}$ denote $\mathcal{K} \times \mathcal{K}$ and $\mathcal{K}^* \times \mathcal{K}^*$ orthogonal matrices, respectively. Write $\mathcal{S}Q_{\mathcal{K}}$ and $\widetilde{\mathcal{S}}Q_{\mathcal{K}^*} \subset \mathbb{R}^p$ to denote the subspaces spanned by the columns $SQ_{\mathcal{K}}$ and $\widetilde{S}Q_{\mathcal{K}^*}$. Then, it is straightforward to verify the invariance of $\mathbb{F}$ to orthogonal transformations of subspace basis vectors:

$$
\mathbb{F}(\mathcal{S}, \mathcal{S}Q_{\mathcal{K}}) = 0, \qquad \mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \mathbb{F}(\mathcal{S}Q_{\mathcal{K}}, \widetilde{\mathcal{S}}Q_{\mathcal{K}^*}).
$$

That is, the $\mathbb{R}^p$-subspaces $\mathcal{S}$ and $\mathcal{S}\mathcal{Q}_\mathcal{K}$ (or $\mathcal{S}\mathcal{Q}_\mathcal{K}$ and $\widetilde{\mathcal{S}}\mathcal{Q}_{\mathcal{K}^*}$) are the same (unaffected by the orthogonal transformation of the columns of the matrix S) which the metric $\mathbb{F}(\cdot,\cdot)$ captures by declaring the distance between $\mathcal{S}$ and $\mathcal{S}\mathcal{Q}_\mathcal{K}$ to be zero.

The above equalities follow directly from the preservation of projection matrices under orthogonal matrix transformations:

$$
\begin{aligned}
P_{\mathcal{S}\mathcal{Q}_\mathcal{K}} &= SQ_\mathcal{K}(Q_\mathcal{K}^\top S^\top SQ_\mathcal{K})^{-1}Q_\mathcal{K}^\top S^\top = SQ_\mathcal{K}[Q_\mathcal{K}^\top(S^\top S)^{-1}Q_\mathcal{K}]Q_\mathcal{K}^\top S^\top \\
&= S(S^\top S)^{-1}S^\top = P_\mathcal{S}.
\end{aligned}
$$

If Q is a $p \times p$ orthogonal matrix and $\mathcal{Q}\mathcal{S}$, $\mathcal{Q}\widetilde{\mathcal{S}}$ denote the column spaces spanned by QS and $Q\widetilde{S}$, it also holds that $\mathbb{F}(\mathcal{S},\widetilde{\mathcal{S}}) = \mathbb{F}(\mathcal{Q}\mathcal{S},\mathcal{Q}\widetilde{\mathcal{S}})$ for invariance to transformations of the row space.

It is important that any metric between subspaces recognizes that orthogonal transformations do not affect the subspace so that this invariance which the metric $\mathbb{F}(\cdot,\cdot)$ possesses is really a minimal condition to require of a subspace metric.

2.5.2 Upper bounds

To explain some further properties of the vector space metric $\mathbb{F}$, we first give an upper bound to the metric.

Lemma 2.1 *Let $\mathcal{S}$, $\widetilde{\mathcal{S}}$ be two vector subspaces of $\mathbb{R}^p$. The Frobenius norm-based metric $\mathbb{F}(\mathcal{S},\widetilde{\mathcal{S}}) = [\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - 2\mathrm{tr}(P_\mathcal{S}P_{\widetilde{\mathcal{S}}})]^{\frac{1}{2}}$ is bounded from above by*

$$
\mathbb{F}(\mathcal{S},\widetilde{\mathcal{S}}) \leq \sqrt{p}.
$$

Proof. Applying (2.2) and Lemma 7.3, we find that

$$
\mathbb{F}(\mathcal{S},\widetilde{\mathcal{S}}) \leq [\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}})]^{\frac{1}{2}}. \tag{2.13}
$$

We now consider two cases to establish Lemma 2.1:

CASE 1: $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) \leq p,$

CASE 2: $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) > p.$

CASE 1: In the first case, Lemma 2.1 follows easily from (2.13).

CASE 2: To handle the second case, note that $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \mathbb{F}(\mathcal{S}^{\perp}, \widetilde{\mathcal{S}}^{\perp})$, which follows from the fact that $P_{\mathcal{S}} + P_{\mathcal{S}^{\perp}} = \mathrm{I}_p = P_{\widetilde{\mathcal{S}}} + P_{\widetilde{\mathcal{S}}^{\perp}}$, where I_p denotes the identity matrix of rank p. Taking traces of the two matrix sums, we find

$$
\begin{aligned}
2p &= \mathrm{tr}(P_{\mathcal{S}}) + \mathrm{tr}(P_{\mathcal{S}^{\perp}}) + \mathrm{tr}(P_{\widetilde{\mathcal{S}}}) + \mathrm{tr}(P_{\widetilde{\mathcal{S}}^{\perp}}) \\
&= \dim(\mathcal{S}) + \dim(\mathcal{S}^{\perp}) + \dim(\widetilde{\mathcal{S}}) + \dim(\widetilde{\mathcal{S}}^{\perp}).
\end{aligned}
$$

Consequently, if $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) > p$, it must be the case that $\dim(\mathcal{S}^{\perp}) + \dim(\widetilde{\mathcal{S}}^{\perp}) < p$. Then, Lemma 2.1 follows easily from $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \mathbb{F}(\mathcal{S}^{\perp}, \widetilde{\mathcal{S}}^{\perp}) = [\dim(\mathcal{S}^{\perp}) + \dim(\widetilde{\mathcal{S}}^{\perp}) - 2\mathrm{tr}(P_{\mathcal{S}^{\perp}} P_{\widetilde{\mathcal{S}}^{\perp}})]^{\frac{1}{2}} \leq \sqrt{p}$ and Lemma 7.3. $\square$

We can show that the metric $\mathbb{F}$ agrees with the intuition that two vector subspaces $\mathcal{S}, \widetilde{\mathcal{S}}$ in $\mathbb{R}^p$ should be maximally distant if one space is the orthogonal complement of the other in $\mathbb{R}^p$ (see also Example 2.1, Section 2.4). The following result establishes this property of $\mathbb{F}$.

Theorem 2.1 *Let $\mathcal{S}, \widetilde{\mathcal{S}}$ be two vector subspaces of $\mathbb{R}^p$. The Frobenius norm-based metric $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}})$ will take on its maximum value $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{p}$ if and only if $\widetilde{\mathcal{S}} = \mathcal{S}^{\perp}$.*

Proof. In the following, we suppose spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ are each spanned by an orthonormal basis $S = [s_1, \ldots, s_{\mathcal{K}}]$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{\mathcal{K}^*}]$, respectively, where $\mathcal{K}, \mathcal{K}^* \leq p$. Write $P_{\mathcal{S}}, P_{\widetilde{\mathcal{S}}}$ and $P_{\mathcal{S}^{\perp}}$ to denote the corresponding projection matrices onto $\mathcal{S}, \widetilde{\mathcal{S}}$ and $\mathcal{S}^{\perp}$, respectively.

If $\widetilde{\mathcal{S}} = \mathcal{S}^{\perp}$, it follows that $P_{\widetilde{\mathcal{S}}} = \mathrm{I}_p - P_{\mathcal{S}} = P_{\mathcal{S}^{\perp}}$ and $P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}} = 0$ and $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) = p$. Applying these facts in (2.2), we find $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{p}$, which is the maximal distance between the vector subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ by Lemma 2.1.

We now suppose $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{p}$ and show that $\widetilde{\mathcal{S}} = \mathcal{S}^\perp$. We can first argue that $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) = p$ holds. That is, if $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) < p$, it would follow that $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) < \sqrt{p}$ by (2.2) and Lemma 7.3, which is a contradiction. Likewise, if $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) > p$, we would again have a contradiction by $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \mathbb{F}(\mathcal{S}^\perp, \widetilde{\mathcal{S}}^\perp) < \sqrt{p}$ by (2.2) and Lemma 7.3.

Consequently, we proceed assuming that $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{p}$ and $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) = p$, from which it follows that $\mathrm{tr}(P_\mathcal{S} P_{\widetilde{\mathcal{S}}}) = 0$ by (2.2). Writing $A = S^\top \widetilde{S} = L \Delta M^\top$ using the singular value decomposition from Lemma 7.2, we can express

$$0 = \mathrm{tr}(P_\mathcal{S} P_{\widetilde{\mathcal{S}}}) = \mathrm{tr}(AA^\top) = \sum_{i=1}^{\mathcal{K}'} \delta_i^2,$$

where $\mathcal{K}'$ represents the rank of $AA^\top$ and $\delta_1^2, \ldots, \delta_{\mathcal{K}'}^2$ are the nonnegative eigenvalues of $AA^\top$. Consequently, each $\delta_i = 0$ for $i = 1, \ldots, \mathcal{K}'$, which implies that the $\mathcal{K} \times \mathcal{K}^*$ matrix $\Delta = 0_{\mathcal{K} \times \mathcal{K}^*}$ has only zero entries and also that $S^\top \widetilde{S} = A = 0_{\mathcal{K} \times \mathcal{K}^*}$. Then it follows that the spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal. By this orthogonality and $\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) = p$, we have $\widetilde{\mathcal{S}} = \mathcal{S}^\perp$. $\square$

In some of the robustness studies to follow, it is also useful to know the maximal possible distance between two vector spaces $\mathcal{S}, \widetilde{\mathcal{S}} \subset \mathbb{R}^p$ w.r.t. the metric $\mathbb{F}$, conditioned on the additional information (restriction) that these spaces have the same size or dimension. We give a modification of Lemma 2.1 to frame the largest value of $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}})$ in this case.

Lemma 2.2 *Suppose $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ are subspaces both of dimension $\mathcal{K} \leq p$. Then the intersection of $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ is at least of dimension $\max\{0, 2\mathcal{K} - p\}$. That is, $\dim(\mathcal{S} \cap \widetilde{\mathcal{S}}) \geq \max\{0, 2\mathcal{K} - p\}$.*

Proof. Define a vector subspace $\mathcal{U} = \{s + \widetilde{s} : s \in \mathcal{S}, \widetilde{s} \in \widetilde{\mathcal{S}}\} = \mathcal{S} \oplus \widetilde{\mathcal{S}} \subset \mathbb{R}^p$. By Theorem 4.4.19 of Meyer, (2000), it holds that

$$\dim(\mathcal{S} \cap \widetilde{\mathcal{S}}) = \dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - \dim(\mathcal{U}) = 2\mathcal{K} - \dim(\mathcal{U}) \geq 2\mathcal{K} - p,$$

using $\dim(\mathcal{U}) \leq p$. The claim now follows. $\square$

Theorem 2.2 *Suppose two vector spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ are each spanned by an orthonormal basis $S = [s_1, \ldots, s_{\mathcal{K}}]$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{\mathcal{K}}]$, $\mathcal{K} \leq p$, so that $\dim(\mathcal{S}) = \dim(\widetilde{\mathcal{S}}) = \mathcal{K}$. Then, the metric based on the Frobenius norm is bounded by*

$$\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \sqrt{2\big(\mathcal{K} - \max\{0, 2\mathcal{K} - p\}\big)}.$$

When $2\mathcal{K} \leq p$, then $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2\mathcal{K}}$ if and only if the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal; when $2\mathcal{K} > p$, then $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2\big(\mathcal{K} - \{2\mathcal{K} - p\}\big)}$ if and only if the subspaces $\mathcal{S} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^\perp$ and $\widetilde{\mathcal{S}} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^\perp$ are orthogonal.

Proof. To establish the claim we need to consider two cases: $2\mathcal{K} \leq p$ and $2\mathcal{K} > p$.

CASE 1. Consider $2\mathcal{K} \leq p$. Since $\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}}) \geq 0$ by Lemma 7.3 and $\mathcal{K} = \mathrm{tr}(P_{\mathcal{S}}) = \mathrm{tr}(P_{\widetilde{\mathcal{S}}})$, it follows that

$$\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \big[\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - 2\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}})\big]^{\frac{1}{2}} \leq \sqrt{2\mathcal{K}},$$

by the definition of $\mathbb{F}$ in (2.2). It also follows easily that $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2\mathcal{K}}$ if and only if $\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}}) = 0$, which is equivalent to the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ being orthogonal to each other by Lemma 7.3.

CASE 2. Assume here $2\mathcal{K} > p$. Let $d = \dim(\mathcal{S} \cap \widetilde{\mathcal{S}})$ and define a $p \times d$ matrix A with orthonormal columns that are basis vectors for the intersection subspace $\mathcal{S} \cap \widetilde{\mathcal{S}}$. If P_A denotes the projection matrix for $\mathcal{S} \cap \widetilde{\mathcal{S}}$, then it follows from Definition 2.6 that $P_{\mathcal{S}} - P_A$ and $P_{\widetilde{\mathcal{S}}} - P_A$ are the projection matrices for the subspaces $\mathcal{S} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^\perp$ and $\widetilde{\mathcal{S}} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^\perp$, respectively (i.e, after removing vectors common to both $\mathcal{S}$ and $\widetilde{\mathcal{S}}$). Then,

$$
\begin{aligned}
\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) &= \big[\dim(\mathcal{S}) + \dim(\widetilde{\mathcal{S}}) - 2\mathrm{tr}(P_{\mathcal{S}} P_{\widetilde{\mathcal{S}}})\big]^{\frac{1}{2}} \\
&= \big[2\mathcal{K} - 2\mathrm{tr}\big([P_{\mathcal{S}} - P_A + P_A][P_{\widetilde{\mathcal{S}}} - P_A + P_A]\big)\big]^{\frac{1}{2}} \\
&= \big[2\mathcal{K} - 2\mathrm{tr}\big([P_{\mathcal{S}} - P_A][P_{\widetilde{\mathcal{S}}} - P_A] + P_A\big)\big]^{\frac{1}{2}} \\
&= \big[2\mathcal{K} - 2d - 2\mathrm{tr}\big([P_{\mathcal{S}} - P_A][P_{\widetilde{\mathcal{S}}} - P_A]\big)\big]^{\frac{1}{2}}, \qquad (2.14)
\end{aligned}
$$

using above that $P_{\mathcal{S}} P_A = P_A = P_{\widetilde{\mathcal{S}}} P_A$, since $\mathcal{S} \cap \widetilde{\mathcal{S}}$ is a subset of both $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, and P_A is idempotent with $d = \mathrm{rank}(A) = \mathrm{tr}(P_A)$. Because $\mathrm{tr}\big([P_{\mathcal{S}} - P_A][P_{\widetilde{\mathcal{S}}} - P_A]\big) \geq 0$ by

Lemma 7.3 (namely, $P_{\mathcal{S}} - P_A$ and $P_{\widetilde{\mathcal{S}}} - P_A$ are the projection matrices) and $d \geq 2\mathcal{K} - p$ by Lemma 2.2, we now find in (2.14) that

$$\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) \leq \sqrt{2\big(\mathcal{K} - \{2\mathcal{K} - p\}\big)}.$$

It also stands from (2.14) that $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2\big(\mathcal{K} - \{2\mathcal{K} - p\}\big)}$ if and only if $\mathrm{tr}\big([P_{\mathcal{S}} - P_A][P_{\widetilde{\mathcal{S}}} - P_A]\big) = 0$ (and $d = 2\mathcal{K} - p$), which is equivalent to the orthogonality of the subspaces $\mathcal{S} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^{\perp}$ and $\widetilde{\mathcal{S}} \cap (\mathcal{S} \cap \widetilde{\mathcal{S}})^{\perp}$ by Lemma 7.3. $\square$

Remark. The distinction between the result above and the one in Lemma 2.1 is that, if $2\mathcal{K} > p$, then the vector spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}} \subset \mathbb{R}^p$ must have a nonempty intersection, namely $\mathcal{I} = \mathcal{S} \cap \widetilde{\mathcal{S}} \neq \emptyset$. In this case, it is clear that $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ cannot be completely orthogonal. However, the proof shows that, when $2\mathcal{K} > p$, the spaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are *most distant* w.r.t. the $\mathbb{F}$ metric when spaces $\mathcal{S} \cap \mathcal{I}^{\perp}$ and $\widetilde{\mathcal{S}} \cap \mathcal{I}^{\perp}$ (after removing any vectors contained in both $\mathcal{S}$ and $\widetilde{\mathcal{S}}$) are orthogonal. In a sense, this agrees with geometric intuition. On the other hand, if $2\mathcal{K} \leq p$, we find that $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are *most distant* w.r.t. the $\mathbb{F}$ metric, namely $\mathbb{F}(\mathcal{S}, \widetilde{\mathcal{S}}) = \sqrt{2\mathcal{K}}$, when these spaces are orthogonal (but $\mathcal{S} \cup \widetilde{\mathcal{S}}$ may not necessarily span all of $\mathbb{R}^p$).

Chapter 3

Sliced Inverse Regression − A Review

Consider the task of estimating a nonparametric regression function g based on observed data points $(x_i, y_i), i = 1, \ldots, n$ of a random vector $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, where Y denotes the response variable and $X = (X_1, \ldots, X_p)^\top$ is a vector of regressor variables. That is, we wish to understand an underlying model

$$Y = g(X, \varepsilon) \tag{3.1}$$

that explains the response Y in terms of the regressors X up to an error term ε which is independent of X. A common difficulty that can arise is the so-called CURSE OF DIMENSIONALITY introduced by Bellman (1961) (see also Friedman (1991), Gather & Becker (2001)) which describes the dilemma that the volume of a p-dimensional space grows exponentially as a function of the dimension p. As an immediate consequence, it becomes more difficult to accurately estimate a function $g(\cdot)$ because more data (x_i, y_i) is needed in order to fill the space $\mathbb{R}^p \times \mathbb{R}$ densely enough to fit the function g.

Fortunately, it often turns out that an appropriate subspace of the original regressor space $\mathbb{R}^p$ is sufficient for an adequate fit of g. Such a subspace can be obtained by

identifying important linear combinations of the regressor variables, $B^\top X$, using a $p \times \mathcal{K}$ matrix of linearly independent column vectors $B = [\beta_1, \ldots, \beta_\mathcal{K}]$ for some $\mathcal{K} < p$, such that the dimension of the essential regressor space is *reduced* from p to $\mathcal{K}$. Li (1991) introduced a procedure for identifying such important linear combinations, i.e. the vectors B, called Sliced Inverse Regression (Sir). The theoretical properties of the procedure have been originally developed by Li and Duan (1987) and Duan and Li (1991). Once the linear combinations of the regressors, $B^\top X$, are identified, these can be used to fit a regression model to predict the response variable Y.

The underlying idea of Sir is to model p one-dimensional regression problems instead of one p-dimensional regression problem and thus avoiding the curse of dimensionality. This is done by considering the inverse regression curve $\mathrm{E}(X|Y)$ rather than the usual curve $\mathrm{E}(Y|X)$ considered in multiple regression. Under some model assumptions, important linear combinations of the regressor variables $B^\top X$ correspond to those which are needed to linearly span and completely explain the inverse regression curve $\mathrm{E}(X|Y)$ in the $\mathbb{R}^p$ space.

In contrast to other dimension reduction procedures, Sir is not able to identify the directions $\beta_1, \ldots, \beta_\mathcal{K}$ itself, which are also not unique, but the $\mathbb{R}^p$-subspace spanned by the directions $\mathcal{B} = \mathrm{span}(\beta_1, \ldots, \beta_\mathcal{K})$. This subspace is unique and $\mathcal{B}$ is called Effective dimension reduction subspace and vectors $\beta_1, \ldots, \beta_\mathcal{K}$ are referred to as effective dimension reduction vectors or directions (henceforth denoted as the e.d.r. subspace and e.d.r. directions, respectively). The identification of the e.d.r. subspace with Sir is possible via a principal component analysis of a particular covariance matrix $V = \mathrm{Cov}[\mathrm{E}(Z|Y)]$ involving the conditional expectation $\mathrm{E}(Z|Y)$, where $Z = \Sigma^{-1/2}(X - \mathrm{E}(X))$ denotes the standardized regressor vector with $\Sigma = \mathrm{Cov}(X)$. By conducting a spectral decomposition of V (principal component analysis), directions with high variability in the random vector $\mathrm{E}(X|Y)$ can be recognized as the eigenvectors of V corresponding to the largest eigenvalues of V. These eigenvectors of V can be used to provide meaningful linear combinations $B = [\beta_1, \ldots, \beta_\mathcal{K}]$ of the regressor variables X. The eigenvectors of V, corresponding to the largest $\mathcal{K}$ eigenvalues,

correspond to the standardized e.d.r. directions, which we will denote by $\eta_1, \ldots, \eta_{\mathcal{K}}$. Undoing the standardization will then yield the actual e.d.r. directions $\beta_1, \ldots, \beta_{\mathcal{K}}$, whereby $\beta_i = \Sigma^{-1/2}\eta_i$ for $i = 1, \ldots, \mathcal{K}$.

In the following section, we will explain the SIR procedure in more detail. Later sections of the dissertation will focus on the robustness properties of SIR against data contamination. It should be noted that, since the development of SIR by Li (1991), other dimension reduction procedures have been proposed for identifying important linear combinations $B = [\beta_1, \ldots, \beta_{\mathcal{K}}]$ of the regressor variables X, or more precisely the space $\mathcal{B} \subset \mathbb{R}^p$ spanned by B. These include, for example, SAVE (Cook, 2000) or MAVE (Xia et al.). However, we shall give here a detailed study of the robustness of SIR.

3.1   Description of the model

To ease the discussion, we give a precise definition of a spectral decomposition of a symmetric matrix such as required in SIR to identify the eigenvalues and eigenvectors of the matrix V.

Definition 3.1 (A normalized spectral decomposition) *A symmetric $p \times p$ real-valued matrix A may be uniquely written as*

$$A = \sum_{i=1}^{p} \lambda_i \eta_i \eta_i^\top, \quad \lambda_i \in \mathbb{R}, \eta_i \in \mathbb{R}^p, \quad \eta_i^\top \eta_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

Unless stated otherwise, we will adopt the convention that η_i, $i = 1, \ldots, p$ will be referred to as the normalized eigenvectors of A; λ_i, $i = 1, \ldots, p$ are said to be the corresponding eigenvalues of A.

The above definition is a version of the spectral decomposition theorem (cf. Schott, 1995, Theorem 4.2). Of course, when a symmetric matrix A is nonnegative definite, it holds that each eigenvalue $\lambda_i \geq 0$ is nonnegative in the above spectral decomposition.

Before continuing to state the Sir procedure we proceed with reviewing necessary assumptions for Sir. We begin with the statistical model that is assumed to describe the relationship between the response variable Y and the p regressor variables $X = (X_1, \ldots, X_p)^\top$.

Model 3.1 *Assume a nonlinear regression function* $f : \mathbb{R}^{\mathcal{K}+1} \to \mathbb{R}$ *of form*

$$Y = f(\beta_1^\top X, \ldots, \beta_{\mathcal{K}}^\top X, \varepsilon) = f(B^\top X, \varepsilon) \tag{3.2}$$

describing the functional relationship between the response variable $Y \in \mathbb{R}$ *and p regressor variables* $X = (X_1, \ldots, X_p)^\top \in \mathbb{R}^p$. *The following assumptions hold for the above model:*

- $\mu = \mathrm{E}(X)$ *denotes the expected value of X and* $\Sigma = \mathrm{Cov}(X)$ *the corresponding covariance matrix. We assume that Σ is positive definite.*

- $\varepsilon \in \mathbb{R}$ *denotes an error term, independent of the regressor variables X.*

- $\beta_1, \ldots, \beta_{\mathcal{K}} \in \mathbb{R}^p$ *are the (linearly independent) e.d.r. directions spanning the e.d.r. subspace* $\mathcal{B} = \mathrm{span}(\beta_1, \ldots, \beta_{\mathcal{K}})$.

We shall note that the functions g in (3.1) and f in Model 3.1 describing the relationship between X and Y and $B^\top X$ and Y, respectively, will typically not be identical. Furthermore, Sir does not attempt to fit the regression model f, it solely estimates the e.d.r. subspace $\mathcal{B}$. Once an estimate for $\mathcal{B}$ is obtained the model fit is attempted in a separate step.

As we have already pointed out Sir is based on the inverse regression curve $\mathrm{E}(X|Y)$ instead of $\mathrm{E}(Y|X)$. Li (1991) showed that under certain conditions, which we will state below, the standardized inverse regression curve $\Sigma^{-1/2}(\mathrm{E}(X|Y) - \mathrm{E}(X))$ falls in the linear subspace spanned by the directions $\Sigma^{1/2}\beta_1, \ldots, \Sigma^{1/2}\beta_{\mathcal{K}}$. The reason for this can be found in the property that the expected value of an arbitrary linear combination of the regressor variables X conditioned on $\beta_1^\top X, \ldots, \beta_{\mathcal{K}}^\top X$ is again linear in the conditional vectors, which Li formalized in the following condition.

Condition 3.1 (*Li, 1991*) *For any* $b \in \mathbb{R}^p$, *the conditional expectation* $\mathrm{E}(b^\top X | \beta_1^\top X, \ldots, \beta_{\mathcal{K}}^\top X)$ *is linear in* $\beta_1^\top X, \ldots, \beta_{\mathcal{K}}^\top X$; *that is, for some constants* $c_0, c_1, \ldots, c_{\mathcal{K}}$, *we have*

$$\mathrm{E}(b^\top X | \beta_1^\top X, \ldots, \beta_{\mathcal{K}}^\top X) = c_0 + c_1 \beta_1^\top X + \ldots + c_{\mathcal{K}} \beta_{\mathcal{K}}^\top X.$$

It can be shown that Condition 3.1 is fulfilled if in the assumed Model 3.1, the random vector $X \in \mathbb{R}^p$ is characterized by a non-degenerate elliptically symmetric distribution; see e.g. Li (1991), Duan and Li (1991) or Hilker (1997). For more detailed discussions on this assumption we refer to Diaconis and Freedman (1984), Eaton (1986) or Cook and Weisberg (1991), Hall and Li (1993) or Kötter (2000).

Condition 3.1 ensures that the centered inverse regression curve $\mathrm{E}(X|Y) - \mathrm{E}(X)$ lies in the linear subspace spanned by $\Sigma\beta_i, i = 1 \ldots, \mathcal{K}$, stated in the next theorem and proved by Li (1991).

Theorem 3.1 (*Li, 1991*) *For the assumed Model 3.1 under Condition 3.1, the centered inverse regression curve* $\mathrm{E}(X|Y) - \mathrm{E}(X)$ *is contained in the linear subspace spanned by* $\Sigma\beta_i, i = 1 \ldots, \mathcal{K}$, *where* $\Sigma = Cov(X)$.

This property also holds true if the centered inverse regression curve $\mathrm{E}(X|Y) - \mathrm{E}(X)$ is standardized, which can be explained as follows. It can be shown that, under Model 3.1 and Condition 3.1, the $p \times p$ covariance matrix $V = \mathrm{Cov}[\Sigma^{-1/2}\mathrm{E}(X|Y)]$ admits the following spectral decomposition

$$V = \mathrm{Cov}[\Sigma^{-1/2}\mathrm{E}(X|Y)] = \sum_{i=1}^{\mathcal{K}} \lambda_i \eta_i \eta_i^\top \tag{3.3}$$

with eigenvalues $0 < \lambda_{\mathcal{K}} \leq \cdots \leq \lambda_1$ and corresponding $\mathcal{K}$ orthonormal eigenvectors $\{\eta_i\}_{i=1}^{\mathcal{K}}$; see Li (1991) and Duan and Li (1991). The eigenvectors $\eta_1, \ldots, \eta_{\mathcal{K}}$, corresponding to the positive eigenvalues of V, are again referred to as the standardized e.d.r. directions. Li (1991) shows that the standardized regression curve $\Sigma^{-1/2}(\mathrm{E}(X|Y) - \mathrm{E}(X))$ is spanned by $\eta_1, \ldots, \eta_{\mathcal{K}}$, as summarized in the next result.

Corollary 3.1 *(**Li, 1991**) Assume that $Z = \Sigma^{-1/2}(X - \mathrm{E}(X))$ has been standardized. Then under the Model 3.1 and Condition 3.1, the standardized inverse regression curve $\mathrm{E}(Z|Y)$ is contained in the linear space generated by the standardized e.d.r. directions $\eta_1, \ldots, \eta_\mathcal{K}$.*

It follows from Theorem 3.1 and Corollary 3.1 that the actual e.d.r. directions are then $\beta_i = \Sigma^{-1/2}\eta_i$, $i = 1, \ldots, \mathcal{K}$, using the normalized eigenvectors (3.3) of V.

As an alternative to the covariance matrix $V = \mathrm{Cov}[\Sigma^{-1/2}\mathrm{E}(X|Y)]$, Li (1991) suggests a covariance matrix approximation based on *slicing*. The range of the response variable is divided into H consecutive and disjoint intervals (slices), denoted by $\mathbb{I}_i$, $i = 1, \ldots, H$. The slices yield sliced proportions $p_i = \mathrm{P}(Y \in \mathbb{I})$ and sliced means $\mu_i = \mathrm{E}(X|Y \in \mathbb{I}_i) \in \mathbb{R}^p$, $i = 1, \ldots, H$, where $\sum_{i=1}^{H} p_i\mu_i = \mu = \mathrm{E}(X)$. Then, the covariance matrix $V = \mathrm{Cov}[\Sigma^{-1/2}\mathrm{E}(X|Y)]$ is replaced with a step-wise approximation

$$V = \Sigma^{-1/2}\sum_{i=1}^{H} p_i(\mu_i - \mu)(\mu_i - \mu)^\top\Sigma^{-1/2}$$

using the same notation V as a small notational abuse. The idea is that this alternative covariance matrix formulation V is often easier to estimate with data.

3.2   Sir with known subspace dimension $\mathcal{K}$

We have now set all the necessary groundwork to give an explicit description of the SIR procedure consisting of the following six steps. Assume we have observed data points $(X, Y)^n = ((x_1, y_1), \ldots, (x_n, y_n))$ with $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$ for $i = 1, \ldots, n$ and the dimension of the e.d.r. subspace $\mathcal{K}$ is known.

1. Standardization of $x_1, \ldots, x_n$ yielding observations

   $$z_i = \widehat{\Sigma}^{-1/2}(x_i - \overline{x}), \quad i = 1, \ldots, n,$$

   where $\overline{x}$ denotes the arithmetic mean and $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})^\top$ the sample covariance matrix.

2. Order observations of the response variable from smallest to largest $y_{(1)}, \ldots, y_{(n)}$ and place the corresponding $z_1, \ldots, z_n$ into H slices $\mathbb{I}_h$ with n_h, $h = 1, \ldots, H$, denoting the number of observations within each slice; namely,

$$y_{(1)}, \ldots, y_{(n_1)} \in \mathbb{I}_1; \; y_{(n_1+1)}, \ldots, y_{(n_1+n_2)} \in \mathbb{I}_2; \; \cdots \; y_{(n_1+\cdots+n_{H-1}+1)}, \ldots, y_{(n)} \in \mathbb{I}_H.$$

3. Computation of the sliced means

$$\widehat{m}_h = \frac{1}{n_h} \sum_{i \, : \, y_i \in \mathbb{I}_h} z_i, \quad h = 1, \ldots, H.$$

4. Computation of the covariance matrix of the sliced means

$$\widehat{V} = \frac{1}{n} \sum_{h=1}^{H} n_h \widehat{m}_h \widehat{m}_h^\top = \sum_{h=1}^{H} \widehat{p}_h \widehat{m}_h \widehat{m}_h^\top,$$

where $\widehat{p}_h$ denotes the slice sample proportion. Conduct a (weighted) principal component analysis yielding ordered eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ with corresponding eigenvectors $\widehat{\eta}_1, \ldots, \widehat{\eta}_p$.

5. The eigenvectors corresponding to the $\mathcal{K}$ largest eigenvalues of $\widehat{V}$ are used to estimate $\mathcal{K}$ e.d.r. directions

$$\widehat{\beta}_i = \widehat{\Sigma}^{-1/2} \widehat{\eta}_i, \quad i = 1, \ldots, \mathcal{K}.$$

6. The estimate of the e.d.r. subspace is

$$\widehat{\mathcal{B}} = \mathrm{span}(\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}).$$

With $\widehat{\mathcal{B}} = \mathrm{span}(\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}})$ in hand, one can attempt to fit a model for f.

Modification for Sir with known covariance structure of X

We also note that there is a possible modification to the above steps in the Sir procedure, depending on what information is available. In particular, if the covariance structure of the regressor variable X is known and $\mathrm{E}(X) = \mu \in \mathbb{R}^p$, $\mathrm{Cov}(X) = \Sigma$ are available, then replace $\bar{x}$ with μ and $\widehat{\Sigma}$ with Σ in Steps 1 and 5 above.

3.3 Sir with unknown subspace dimension $\mathcal{K}$

In most practical applications of Sir, the dimension $\mathcal{K}$ of the dimension reduction subspace is unknown and must be estimated from the data $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n$. That is, the data must be used to first find an estimate $\widehat{\mathcal{K}}$ of $\mathcal{K}$ and then obtain an estimated e.d.r. subspace $\widehat{B}_{\widehat{\mathcal{K}}} \subset \mathbb{R}^p$ spanned by the first $\widehat{\mathcal{K}}$ e.d.r. directions, namely $\widehat{B}_{\widehat{\mathcal{K}}} = \mathrm{span}(\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{\mathcal{K}}})$.

Assuming that the regressor variables $X \in \mathbb{R}^p$ follow a multivariate normal distribution ($X \sim \mathcal{N}(\mu, \Sigma)$), Li (1991) suggests the following procedure for estimating $\mathcal{K}$. This method involves performing consecutive hypothesis tests based on the ordered eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ from the Sir procedure, where the covariance structure μ, Σ of X maybe known or unknown. For $0 < \alpha < 1$, let $\chi^2_{\nu,\alpha}$ denote the upper α quantile of a chi-square distribution χ^2_ν with ν degrees of freedom; that is $\mathrm{P}(\chi^2_\nu \geq \chi^2_{\nu,\alpha}) = \alpha$.

Li (1991)'s procedure for estimating $\mathcal{K}$:

1. Pick a level of significance α and set $j = 0$.

2. Using the test statistic $t_j = n \sum_{i=j+1}^p \widehat{\lambda}_i$, test the hypotheses $H_0^j : \mathcal{K} = j$ versus $H_1^j : \mathcal{K} > j$ by rejecting H_0^j if $t_j > \chi^2_{(p-j)(H-j-1),\alpha}$.

3. If H_0^j is rejected, set $j = j + 1$ and repeat Step 2; otherwise set $\widehat{\mathcal{K}} = j$.

Hence, the estimated dimension $\widehat{\mathcal{K}}$ of the e.d.r. subspace is determined by the eigenvalues that are *significantly large*. The estimate $\widehat{\mathcal{K}}$ of $\mathcal{K}$ must be substituted in Steps 5 and 6 of the Sir procedure in Section 3.2. Alternative methods for estimating $\mathcal{K}$ in Sir have been proposed by Ferré (1998), Schott (1994), or Velilla (1998), for example. Schott proposes a test procedure that relaxes Li's assumption of a normally distributed regressor vector X to X following an elliptically symmetric distribution. But all methods for estimating $\mathcal{K}$ involve considerations of the eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ of $\widehat{V}$ from Step 4 in the Sir procedure. More detailed information on these and further methods

can be found in Holland-Letz (2001) who carried out a comparative study assessing their performance with respect to the correct determination of $\mathcal{K}$.

As the fit of a model for f heavily relies on the subspace estimate $\widehat{\mathcal{B}}$, the robustness of the SIR procedure is of crucial interest. In particular it appears decisive to assess the influence of one or more observations $(x_i, y_i), i = 1, \ldots, n$ on the estimate $\widehat{\mathcal{B}}$ of the e.d.r. subspace. We will approach this task in the next chapter.

Chapter 4

A Sensitivity Analysis of Sliced Inverse Regression

Throughout this chapter we wish to investigate the sensitivity (robustness) of the SIR procedure with respect to contamination of the data. We suppose that SIR produces an estimate $\widehat{\mathcal{B}}$ of the e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$ based on a sample $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ of size n. We then wish study the robustness of SIR when one or more observations in $(X, Y)^n$ are replaced with maliciously contaminated values, producing a contaminated subspace estimate $\widehat{\mathcal{B}}^*$. In order to assess this influence of contamination, we quantify the discrepancy of $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}^*$ based on a suitable metric. Consequently, the case we are treating here is the finite sample case.

We begin with a review of previous work in this area in Section 4.1.1, which can mainly be attributed to Hilker (1997), Becker (2001) and Gather et al. (2002). A discussion clarifying what we regard as complete failure of the SIR procedure shall then be considered. We will then continue with a summary of the most important and interesting results and discuss them in more detail with supporting proofs and arguments throughout the rest of Chapter 4.

4.1   Breakdown of Sir

Sir pursuits a dimension reduction of a p-dimensional regressor space by producing a $\mathcal{K}$-dimensional ($\mathcal{K} < p$) subspace estimate $\widehat{\mathcal{B}}$ without any loss of information regarding the functional relationship of the response variable Y with the p regressors $X = (X_1, \ldots, X_p)^\top$. Hence, the parameter of interest with Sir (and other dimension reduction procedures like Save (cf. Cook, 2000)) is a *subspace* of $\mathbb{R}^p$. In the following we will define any failure of Sir with respect to the final subspace estimate and not with respect to any of the individual estimators and respective estimates that are integrated into the steps of the Sir procedure, as done in a first approach by Hilker (1997).

4.1.1   Review of previous breakdown definitions

Hilker (1997) was the first to approach the problem of defining a finite sample breakdown point for Sliced Inverse Regression. He established a definition which is based on the idea of considering the breakdown behavior of all individual functionals the Sir procedure consists of. As seen in Chapter 3, there are four different functionals involved when estimating the e.d.r. subspace $\mathcal{B}$: two location functionals, say $\mathcal{T}_1$ and $\mathcal{T}_2$, as well as two scatter functionals, say $\mathcal{C}_1$ and $\mathcal{C}_2$. Functionals $\mathcal{T}_1$ and $\mathcal{C}_1$ are used in the standardization step of the procedure, while $\mathcal{T}_2$ is used for estimating the sliced means in Step 3 and $\mathcal{C}_2$ for estimating the covariance matrix of the sliced means prior the principal component analysis. In the Sir procedure in Chapter 3, the original location and scatter functionals $\mathcal{T}_1$, $\mathcal{C}_1$, $\mathcal{T}_2$ and $\mathcal{C}_2$ applied to the data correspond to the sample mean $\overline{x}$, sample covariance $\widehat{\Sigma}$, sliced sample means $\widehat{m}_h$, and sample covariance $\widehat{V}$ of the sliced sample means.

Among Hilker's findings is the important result (see also Gather et al. (2001)) that the location functional $\mathcal{T}_1$ in the standardization step does not influence the estimation of the e.d.r. subspace regardless of any potential contamination. For this reason, Hilker argued that focusing on the remaining functionals $\mathcal{T}_2$, $\mathcal{C}_1$ and $\mathcal{C}_2$ is sufficient when

studying the breakdown behavior of SIR. The findings by Hilker can be summarized as follows (see also Becker (2001)). Hilker characterized the breakdown behavior of SIR by trying to identify the *weakest link* in terms of robustness among the individual location and scatter functionals $\mathcal{T}_2$, $\mathcal{C}_1$ and $\mathcal{C}_2$ required in the steps of SIR. It is well-known that many functionals based on sample means and covariances are notoriously non-robust. Hilker (1997) formulated that *breakdown* of SIR could be caused by the breakdown of a scatter functional $\mathcal{C}_1$ in the standardization step, which happens when either the largest eigenvalue of the covariance estimate of Σ (produced by $\mathcal{C}_1$) converges to infinity or the smallest eigenvalue converges to zero [cf. Lopuhaä and Rousseeuw (1991)]. A detailed discussion of this can be found in Hilker's dissertation (p. 93 ff.) but also in Becker (2001, p. 53). Regardless of the behavior of $\mathcal{C}_1$, Hilker suggested that SIR could also break down if the scatter functional $\mathcal{C}_2$ breaks down, which happens when the largest eigenvalue of the covariance matrix estimate of the sliced means, obtained from $\mathcal{C}_2$, becomes arbitrarily large; see Becker (2001, p. 50). Note that the location functional $\mathcal{T}_2$ also contributes to the covariance estimate based on $\mathcal{C}_2$ and is also assessed in terms of the breakdown. Finally, Hilker formulated that a third potential incidence of breakdown could occur when the largest eigenvalue of $\mathcal{C}_2$-based covariance estimate becomes arbitrarily close to zero. Considering the smallest eigenvalue of $\mathcal{C}_2$ is not meaningful, because, due to the nature of the procedure, we can have less than $H < p$ slices resulting in the smallest eigenvalue being equal to zero by default; see also Hilker (1997) and Becker (2001).

Hilker's breakdown definition is based on a combination of the breakdown causes described above, where he defined the breakdown point of SIR as the minimum of the three breakdown points associated with the above breakdown cases (Hilker (1997), Definition 5.10). Becker (2001) advances this definition by recognizing, that following the philosophy of Stromberg and Ruppert (1992), it would probably be more meaningful to look at "the performance of the procedure as a whole." By this, Becker proposes that the definition of breakdown should be formulated with respect to the final product of the SIR procedure, namely the estimated *subspace* itself, rather than looking at breakdown in terms of individual estimators involved in the steps of SIR. According

to Becker, breakdown should more naturally correspond to the situation in which the "subspace is estimated completely wrongly." Although this approach is also hinted at by Hilker (1997), he did not pursue it.

Up to this point, we have not mentioned the distance measures on e.d.r. directions, which both Hilker (1997) and Becker (2001) address. To measure the *distance* between two sets of estimates of e.d.r. directions, based on uncontaminated and contaminated data, both Hilker and Becker proposed using canonical correlations, as originally introduced by Li (1991) to assess the closeness of e.d.r. directions. In particular, when the reduction subspace dimension $\mathcal{K}$ is known, they suggested examining the smallest canonical correlation between uncontaminated estimates of the e.d.r. directions and estimates based on (k-replacement) contaminated data; see Hilker (1997). In fact, Becker gives a first definition of the "correlation breakdown point" based on this smallest canonical correlation when $\mathcal{K}$ is known; see Becker (2001), Definition 4.13. If the smallest canonical correlation becomes zero this implies that at least one estimated e.d.r. direction under contamination must be orthogonal to the e.d.r. subspace estimated without contamination and this is what she defines to be the finite sample breakdown of the Sir procedure. Becker (2001) later extends her finite sample breakdown definition to reflect estimation of the entire e.d.r. subspace under consideration (not just e.d.r. directions), which will shall discuss in the next section.

4.1.2   Issues in formulating failure for Sir

As already noted, a fair discussion regarding the robustness of the Sir procedure should be preceded by clarifying first how the procedure can fail. In order to do so, it helps to first discuss some situations that could possibly represent an erroneous estimate $\widehat{\mathcal{B}}$ of the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$. This task, however, is not as straightforward as it may first appear. Connected to the difficulty of defining erroneous subspace estimates, a further complication is that we require a means to help judge how far off a subspace estimate $\widehat{\mathcal{B}}$ might be from the true e.d.r. subspace $\mathcal{B}$. Of course, one way to assess the

discrepancy or distance between $\mathcal{B}$ and an estimate $\widehat{\mathcal{B}}$ (or between two estimates of $\mathcal{B}$ based on uncontaminated and contaminated data) is through a metric D defined on $\mathbb{R}^p$-subspaces. However, it can be hard to reach a common consensus on a meaningful metric to use.

Listed below, we consider possible erroneous estimates $\widehat{\mathcal{B}}$ for the e.d.r. subspace $\mathcal{B}$ and discuss issues surrounding these. By $\mathcal{K}$, we denote the dimension of the e.d.r. subspace $\mathcal{B}$.

Worst case subspace estimation, $\mathcal{K}$ known

Suppose the dimension $\mathcal{K}$ of the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$ is known. In the extreme cases, if $\mathcal{K}$ is known to be 0 or p, there can be no problems in estimating the reduction subspace, since then either $\mathcal{B} = 0_p$ or $\mathbb{R}^p$ holds. But suppose that $1 \leq \mathcal{K} < p$ is known and $\widehat{\mathcal{B}}$ represents an estimate also with $\dim(\widehat{\mathcal{B}}) = \mathcal{K}$. What type of estimate $\widehat{\mathcal{B}}$ should be considered as the worst case estimate of $\mathcal{B}$ possible? Becker (2001, p. 55) suggests the following "extreme position" as the worst case estimate. She essentially states that we could consider a $\mathcal{K}$-dimensional subspace estimate $\widehat{\mathcal{B}}$ as being extremely distant from the true e.d.r. subspace $\mathcal{B}$ if the following two conditions hold:

C1. the dimension $d = \dim(\mathcal{I})$ of the intersection $\mathcal{I} = \widehat{\mathcal{B}} \cap \mathcal{B}$ is minimal, namely $d = \max\{0, 2\mathcal{K} - p\}$ (see Lemma 2.2);

C2. after removing the intersection $\mathcal{I}$, the subspaces $\widehat{\mathcal{B}} \cap \mathcal{I}^\perp$ and $\mathcal{B} \cap \mathcal{I}^\perp$ are *orthogonal,* where $\mathcal{I}^\perp$ represents the orthogonal complement of $\mathcal{I}$.

(Technically, Becker (2001) considers the sets $\widehat{\mathcal{B}} \setminus \mathcal{I}$ and $\mathcal{B} \setminus \mathcal{I}$, not $\widehat{\mathcal{B}} \cap \mathcal{I}^\perp$ and $\mathcal{B} \cap \mathcal{I}^\perp$, but the first two sets are not $\mathbb{R}^p$-subspaces, e.g., these do not contain 0_p.) Note that this notion of *maximal distance* from Becker *agrees completely* with the mathematical formulation of maximal distance between two $\mathcal{K}$-dimensional subspaces of $\mathbb{R}^p$ based on the Frobenius norm-based subspace metric $\mathbb{F}$ in Definition 2.9. By Theorem 2.2, $\mathcal{K}$-dimensional subspaces $\mathcal{B}$ and $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ are maximally distant with respect to the metric $\mathbb{F}$ if $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) = \sqrt{2(\mathcal{K} - \max\{0, 2\mathcal{K} - p\})}$ which is equivalent to the two conditions

listed above from Becker (2001). Hence, we tend to agree with Becker's concept of worst case estimation of $\mathcal{B}$ when $\mathcal{K}$ is known since this agrees with both geometrical intuition and a mathematically precise metric on subspaces.

Worst case subspace estimation, $\mathcal{K}$ unknown

Suppose now that the dimension $\mathcal{K}$ of $\mathcal{B}$ is unknown and must be estimated, $\widehat{\mathcal{K}}$.

1. Obviously, a subspace estimate $\widehat{\mathcal{B}}$ of $\mathcal{B}$ will be wrong if $\widehat{\mathcal{K}} \neq \mathcal{K}$. But what would be a worst case estimate? When estimating an e.d.r. subspace, a question arising is, whether to equally treat the underestimation and overestimation of $\mathcal{K}$. If one wishes to define breakdown of SIR through a metric to assess distances between $\mathbb{R}^p$-subspaces, then underestimation of $\mathcal{K}$ should, in many cases, be just as erroneous as overestimation, because a metric must be symmetric. On the other hand, one would expect that overestimating $\mathcal{K}$ may not be as bad as underestimation if the resulting subspace estimate $\widehat{\mathcal{B}}$ contains the true reduction subspace $\mathcal{B}$ (i.e., $\mathcal{B} \subset \widehat{\mathcal{B}}$) because overestimation of $\mathcal{K}$ does not cause any loss of information. At the same time, overestimation incorporates some false information in the sense that parts of the orthogonal complement of the e.d.r. subspace $\mathcal{B}$ might be included in the resulting estimate $\widehat{\mathcal{B}}$. In the end, the issue of whether overestimation or underestimation of $\mathcal{K}$ is negative for SIR can only be answered by examining the consequences for estimating the link function f. However, the SIR procedure itself is really only focused on estimating the e.d.r. subspace $\mathcal{B}$ itself and we would like to investigate the robustness properties of SIR independent of fitting the link function f.

2. In the extreme cases that the estimated subspace dimension $\widehat{\mathcal{K}} = 0$ or p, can we consider this as breakdown? Becker (2001, p. 57) suggests that this is one possibility, especially since we would expect the true subspace dimension should satisfy $\mathcal{K} < p$ if dimension reduction is at all possible. However, it is conceivable, though extreme, that true subspace dimension satisfies $\mathcal{K} = 0$ or p, if there is no relationship between the regressors and the response or if no dimension reduction is possible. In this case, estimates $\widehat{\mathcal{K}} = 0$ or p might be allowable.

3. As Becker (2001, p.58) discusses, a formulation of a worst case estimate $\widehat{\mathcal{B}}$ of $\mathcal{B}$, where the dimension $\dim(\widehat{\mathcal{B}}) = \widehat{\mathcal{K}}$ is estimated, should reflect what vectors $\widehat{\mathcal{B}}$ and $\mathcal{B}$ have in common. She suggests using canonical correlations for this. It is geometrically clear that two $\mathbb{R}^p$-subspaces $\widehat{\mathcal{B}}$ and $\mathcal{B}$ should have little in common, or be most distant, if these subspaces are orthogonal. The most extreme case of distance would involve $\widehat{\mathcal{B}} = \mathcal{B}^\perp$, so that $\widehat{\mathcal{B}}$ is the orthogonal complement of $\mathcal{B}$ and contains essentially all $\mathbb{R}^p$-vectors that cannot be spanned in any part by vectors in $\mathcal{B}$. In fact, we could consider using the Frobenius subspace metric to capture this notion. By Theorem 2.1, we find that two subspaces $\widehat{\mathcal{B}}$ and $\mathcal{B}$ of $\mathbb{R}^p$ are maximally distant in terms of the metric $\mathbb{F}$, i.e., $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) = \sqrt{p}$, if and only if $\widehat{\mathcal{B}} = \mathcal{B}^\perp$. Hence, in the case that $\mathcal{K}$ is unknown, we could define extreme failure occurring whenever the estimate $\widehat{\mathcal{B}}$ consists of the orthogonal complement of $\mathcal{B}$ and essentially contains the largest amount of subspace information that is contrary to $\mathcal{B}$.

4. It should not be forgotten that the estimated eigenvalues of $\widehat{V}$ in step 4 of the SIR procedure (see Section 3.2) play an important role in defining worst case subspace estimates. The relative sizes of these estimated eigenvalues determine the estimated e.d.r directions and hence the subspace estimate $\widehat{\mathcal{B}}$. In addition, as discussed in Section 3.3, estimation of $\mathcal{K}$ is often based on the estimated eigenvalues of $\widehat{V}$. Davies and Gather (2004, 2005b) remind us that eigenvalues are central in formulating the robustness of principal component analysis, which SIR involves. Indeed, the studies of the effect of data contamination on SIR, given in Sections 4.2-4.4, always begin with considering the eigenvalues of $\widehat{V}$ under contamination. It is possible as well to formulate breakdown in terms of the estimated eigenvalues of $V = \mathrm{Cov}[\mathrm{E}(\Sigma^{-1/2}X|Y)]$. For example, if all eigenvalues of the estimate $\widehat{V}$ of V are equal, then it becomes impossible to estimate the most influential e.d.r directions needed for any estimate $\widehat{\mathcal{B}}$. We could consider this situation to be a breakdown of SIR as well. This is essentially a point made by Davies and Gather (2004, 2005b).

Again, *breakdown* in the area of robust statistics is commonly understood as *worst case behavior* of a statistical procedure in the sense of infinite bias. From the suggestions of Becker (2001) and our understanding of the subspace metric $\mathbb{F}$ from Definition 2.9, when $\mathcal{K}$ is known a worst case behavior of SIR can be linked to the estimation of a maximally distant subspace $\widehat{\mathcal{B}}$ from the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$ with respect to the metric $\mathbb{F}$, namely $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) = \sqrt{2(\mathcal{K} - \max\{0, 2\mathcal{K} - p\})}$. This again implies that two $\mathcal{K}$-dimensional subspaces $\mathcal{B}$ and $\widehat{\mathcal{B}}$ have little in common through an intersection $\mathcal{I}$ of the smallest possible dimension $\dim(\mathcal{I}) = \max\{0, 2\mathcal{K} - p\}$ and remaining subspace portions $\mathcal{B} \cap \mathcal{I}^\perp, \widehat{\mathcal{B}} \cap \mathcal{I}^\perp$ that are orthogonal.

When $\mathcal{K}$ is unknown, the worst case behavior of SIR appears to be geometrically linked to the estimation of the orthogonal complement $\widehat{\mathcal{B}} = \mathcal{B}^\perp$ of the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$. We can use the metric $\mathbb{F}$ from Definition 2.9 to reflect this because the metric assumes its maximal value $\mathbb{F}(\mathcal{B}, \widehat{\mathcal{B}}) = \sqrt{p}$ when the orthogonal complement of $\mathcal{B}$ is estimated.

This should lead us to the thought that in the finite sample case, when we consider contaminating a data set $(X, Y)^n$ by replacing certain observations with corrupted values, failure of the procedure can considered as *estimation of the orthogonal complement*. More precisely, we could say a subspace estimate based on contaminated data differs maximally from an uncontaminated estimate if the two subspaces have a minimal intersection $\mathcal{I}$ (of dimension $\max\{0, 2\mathcal{K} - p\}$ when $\mathcal{K}$ is known and of dimension 0 when $\mathcal{K}$ is unknown) and non-interesting portions of both subspaces that are orthogonal. This formulation appears to geometrically capture the worst case estimation scenario in both cases where the true subspace dimension $\mathcal{K}$ may be known or unknown. In particular, whether $\mathcal{K}$ is known or unknown, the metric $\mathbb{F}$ from Definition 2.9 provides maximal distances between subspaces that are consistent with this idea of *estimation of the orthogonal complement* as illustrated in Section 2.5.2. Also, $\mathbb{F}$ has the advantage over using single canonical correlations to measure subspace distance, because it constitutes a true metric.

4.2   Discussion of main results

Throughout Sections 4.3 and 4.4 we will investigate in detail the sensitivity of SIR against contaminated data. We will look at the type of contamination that is necessary in order to obtain erroneous (orthogonal) estimates. Already in Hilker (1997), Gather et al. (2002) and Prendergast (2004), it can be found that the influence of a contaminated data point highly depends on the direction in which it is placed.

Mainly, there are two factors that influence a subspace estimate $\widehat{\mathcal{B}}$ of the true e.d.r. subspace $\mathcal{B} \subset \mathbb{R}^p$.

1. *knowledge of the dimension $\mathcal{K}$ of the e.d.r. subspace $\mathcal{B}$*: The relevance of this issue is obvious. If $\mathcal{K}$ is known, the procedure will not under- or overestimate the dimension of $\mathcal{B}$. In the vast majority of all applications however, $\mathcal{K}$ will be unknown and thus constitutes the more realistic case.

2. *knowledge of the covariance structure of the regressor variables* $\mathrm{Cov}(X) = \Sigma$ *and* $\mathrm{E}(X) = \mu$: As indicated in the findings of Hilker (1997) and Gather et al. (2002), contamination schemes that cause damaging subspace estimates in SIR can change depending on whether $\mathrm{Cov}(X) = \Sigma$ and $\mathrm{E}(X) = \mu$ are known or not.

For this reason we will distinguish the following cases in our robustness study and we summarize some of our main findings in Sections 4.3 and 4.4. Suppose $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ denotes a subspace estimate of $\mathcal{B}$ based on a sample $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$, obtained from applying SIR with H slices $(\min\{\mathcal{K}, p\} \leq H)$. We denote an estimate $\widehat{\mathcal{K}}$ of $\mathcal{K} = \dim(\mathcal{B})$ based on $(X, Y)^n$.

- <u>Case I:</u> $\mathcal{K}$, Σ (μ) are known.

  If $1 \leq \mathcal{K} < p$ is known, we can replace $k = \min\{\mathcal{K}, p - \mathcal{K}\}$ observations in $(X, Y)^n$ to obtain a contaminated subspace estimate that is maximally distant (orthogonal) from $\widehat{\mathcal{B}}$ in terms of the Frobenius norm-based subspace metric $\mathbb{F}$, given that

$\mathcal{K}$ is known. The contamination scheme involves replacing a regressor variable x_i with a contaminated value $\widetilde{x}_i = t_m \widetilde{\beta}_i$ across k *different slices*, i.e., one contaminated x-value per slice. To define the k-contaminated x-values, we use a scaling factor $t_m > 0$ determining the magnitude of contamination and k-directions of contamination $\{\widetilde{\beta}_i\}_{i=1}^k$ given by $\widetilde{\beta}_i = \Sigma^{1/2} \overline{\beta}_i$, where $\{\overline{\beta}_i\}_{i=1}^k$ are orthonormal vectors in $\{\Sigma^{1/2} v : v \in \widehat{\mathcal{B}}^{\perp}\}$. In the case $\Sigma = I_p$, the $p \times p$ identity matrix, then the directions of contamination are orthogonal to the uncontaminated subspace estimate $\widehat{\mathcal{B}}$. See Section 4.3.1.

- <u>Case II:</u> $\mathcal{K}$ is known, Σ (μ) is unknown.

If $1 \le \mathcal{K} < p$ is known, the estimate $\widehat{\mathcal{B}}$ from the uncontaminated data $(X, Y)^n$ is spanned by uncontaminated estimates $\{\widehat{\beta}_i\}_{i=1}^{\mathcal{K}}$ of the first $\mathcal{K}$ e.d.r. directions. Again, we can replace $k = \min\{\mathcal{K}, p - \mathcal{K}\}$ observations in $(X, Y)^n$ to obtain a contaminated subspace estimate that is maximally distant (orthogonal) from $\widehat{\mathcal{B}}$ in terms of the Frobenius norm-based subspace metric $\mathbb{F}$, given that $\mathcal{K}$ is known. We use the same contamination scheme as above (i.e., k different slices) with the exception that the k directions of contamination $\{\widetilde{\beta}_i\}_{i=1}^k$ must be a size k subcollection of the $\mathcal{K}$ uncontaminated e.d.r. direction estimates $\{\widehat{\beta}_i\}_{i=1}^{\mathcal{K}}$. That is, contamination is in the direction of the uncontaminated estimates $\{\widehat{\beta}_i\}_{i=1}^{\mathcal{K}}$ when Σ is unknown. See Section 4.3.2.

- <u>Case III:</u> $\mathcal{K}$ is unknown, Σ (μ) is known.

Suppose the uncontaminated estimate of $\mathcal{K}$ satisfies $1 \le \widehat{\mathcal{K}} < p$. We can replace $k = p - \widehat{\mathcal{K}}$ observations in $(X, Y)^n$ so that, under this contamination, the largest k contaminated eigenvalues computed in Step 4 of the SIR procedure explode in size, while the remaining $p - k$ contaminated eigenvalues are bounded. Because the contaminated estimate of the subspace dimension depends on the largest contaminated eigenvalues, we can force the contaminated subspace estimate to have dimension $k = p - \widehat{\mathcal{K}}$ as well as force this subspace estimate to be orthogonal to the uncontaminated $\widehat{\mathcal{B}}$. That is, by replacing $k = p - \widehat{\mathcal{K}}$ observations in $(X, Y)^n$, we can obtain a contaminated subspace estimate that is the orthogonal

complement of $\widehat{\mathcal{B}}$ and so maximally distant (orthogonal) from $\widehat{\mathcal{B}}$ in terms of the Frobenius subspace metric $\mathbb{F}$. The contamination scheme is essentially the same as in CASE I above, replacing x-observations in $k = p - \widehat{\mathcal{K}}$ different slices and using the same contamination directions. See Section 4.4.1.

- CASE IV: $\mathcal{K}$, Σ (μ) are unknown.

  When the uncontaminated dimension estimate is $1 \leq \widehat{\mathcal{K}} < p$, we can replace $k = \widehat{\mathcal{K}}$ observations in $(X, Y)^n$ so that the resulting contaminated subspace estimate is a subspace of $\widehat{\mathcal{B}}^{\perp}$, the orthogonal complement of the uncontaminated estimate $\widehat{\mathcal{B}}$. The contamination scheme is essentially the same as in CASE II above, replacing x-observations in $k = \widehat{\mathcal{K}}$ different slices. But for the k directions of contamination, we use all $k = \widehat{\mathcal{K}}$ uncontaminated e.d.r. directions $\{\widehat{\beta_i}\}_{i=1}^{\widehat{\mathcal{K}}}$ used to span $\widehat{\mathcal{B}}$. Under this contamination, it is very difficult to control the sizes of the contaminated eigenvalue estimates in the SIR procedure (unlike CASE III). Hence, the contaminated estimate of subspace dimension is problematic to directly manage. However, it is possible to show that the contaminated subspace estimate must be orthogonal to $\widehat{\mathcal{B}}$ and must then have dimension less than or equal to $p - \widehat{\mathcal{K}}$. In essence, the contaminated subspace estimate has nothing in common with $\widehat{\mathcal{B}}$ (i.e., is orthogonal), but may not correspond to the entire orthogonal complement of $\widehat{\mathcal{B}}$. See Section 4.4.2.

As in finite sample studies of Hilker (1997) and Gather et al. (2002), we find that directions of contamination that are harmful to SIR differ depending on whether $\mathrm{Cov}(X) = \Sigma$ is known or not. Contamination in the direction of the uncontaminated e.d.r. direction estimates $\widehat{\beta_i}$ is harmful when Σ is unknown, whereas contamination orthogonal to uncontaminated direction estimates is worst when Σ is known. Indeed, the sensitivity of SIR depends on *both* the knowledge of the covariance structure Σ and of the dimension $\mathcal{K}$ of the true e.d.r. subspace $\mathcal{B}$. Hence, our work in the following Sections 4.3-4.4 fully supports and extends the results of Hilker (1997) and Gather et al. (2002).

We finally note that, according to Cook and Critchley (2000), outliers present in a

data set can only cause an overestimation of the number of directions of the subspace in SIR, which supposedly should not be dangerous in estimating the key ingredients of the true e.d.r. subspace. In Sections 4.3 and 4.4, we are able to show that this is not the case.

Throughout the remainder of this Chapter, we assume that the SIR procedure can be applied to a given uncontaminated data set $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n \in \mathbb{R}^p \times \mathbb{R}$ under consideration. Essentially, we suppose that H slices can be identified from the data $(X, Y)^n \in \mathbb{R}^p \times \mathbb{R}$, which always holds true if the y-observations are all distinct.

4.3 Sensitivity Analysis when $\mathcal{K}$ is known

In the case that $1 \leq \mathcal{K} < p$ is known, the estimated reduction subspace $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ computed from the data $(X, Y)^n = \{(x_i, y_i)\}_{i=1}^n$ has dimension $\mathcal{K}$. Suppose that we replace $k \leq n$ arbitrary data points in $(X, Y)^n$ to obtain a contaminated data sample $(X, Y)^{n,k}$. Applying the SIR-procedure with $(X, Y)^{n,k}$ given the known dimension $\mathcal{K}$, we would produce a contaminated estimate of the reduction subspace $\widehat{\mathcal{B}}_k$ spanned by $\mathcal{K}$ independent vectors based on $(X, Y)^{n,k}$.

To judge the effect of data contamination on dimension reduction subspace estimation, we consider measuring the distances between $\mathcal{K}$-dimensional subspaces $\widehat{\mathcal{B}}$ (estimated e.d.r. subspace using uncontaminated data $(X, Y)^n$) and $\widehat{\mathcal{B}}_k$ (e.d.r. subspace estimated by contaminating k data points $(X, Y)^{n,k}$) with the metric $\mathbb{F}$. The worst-case contamination scenario would cause the subspaces $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_k$ to be most distant under the measure $\mathbb{F}$.

When $\mathcal{K}$ is known, we can quantify the worst-case scenario under contamination, for dimension reduction as

$$\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_k) = \sqrt{2\big(\mathcal{K} - \max\{0, 2\mathcal{K} - p\}\big)} = \begin{cases} \sqrt{2\mathcal{K}} & \text{if } 2\mathcal{K} \leq p, \\ \sqrt{2(p - \mathcal{K})} & \text{if } 2\mathcal{K} > p, \end{cases} \tag{4.1}$$

using Theorem 2.2. That is, $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_k$ are maximally distant in terms of the subspace metric $\mathbb{F}$.

We are now prepared to define a finite sample breakdown point for SIR-type dimension reduction procedures, in the case that the dimension $\mathcal{K}$ of reduction is known.

Definition 4.1 *Finite sample breakdown point in dimension reduction ($\mathcal{K}$* *known).* *Let $(X,Y)^{n,k}$ denote a contaminated sample found by replacing $1 \leq k \leq$* *n data points in a data set $(X,Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ with arbitrary values* *$\{(\widetilde{x}_{i_j}, \widetilde{y}_{i_j})\}_{j=1}^k$. For fixed $1 \leq \mathcal{K} < p$, let $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_k \subset \mathbb{R}^p$ denote estimates of $\mathcal{K}$-* *dimensional e.d.r subspaces based on a given dimension reduction procedure (e.g., SIR)* *applied to $(X,Y)^n$ and $(X,Y)^{n,k}$, respectively. The **finite sample breakdown point*** *of the dimension reduction procedure is defined as*

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

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

under the metric $\mathbb{F}$ for the data constellation $(X,Y)^n$.

The value $\epsilon_{fsbp,\mathcal{K}}$ represents the percentage of contamination in a data set $(X,Y)^n$ necessary to cause a dimension reduction procedure to *breakdown* with a maximal distance (4.1) between the subspace estimate under contamination and the estimate produced from an original, uncontaminated sample. Note that in the cases that $\mathcal{K} = 0$ or $\mathcal{K} = p$ is known, the finite sample breakdown point definition above is not applicable because *breakdown* in these extreme cases is not even possible; when $\mathcal{K} = 0$ or $\mathcal{K} = p$, it is known that the e.d.r. subspace estimate should be $\widehat{\mathcal{B}} = 0_p$ or $\widehat{\mathcal{B}} = \mathbb{R}^p$, respectively.

4.3.1 $\mathcal{K}$ known $-\ \Sigma$ known case

The following lemma is useful for determining the limiting behavior of eigenvectors corresponding to a sequence of non-negative definite covariance matrices. Since the estimated reduction space in SIR corresponds to (scaled) estimated eigenvectors, we shall exploit the result to understand the influence of data contamination on SIR in the case of a known covariance structure. We note that Lemma 4.8 of Hilker (1997) follows as a special case of Lemma 4.1 here (i.e., $\mathcal{K} = 1$), but our result is geometrically stronger and more comprehensive in terms of the convergence of subspaces spanned by convergent eigenvectors.

In the following lemma, we use $|A|$ to denote the cardinality of a finite set A.

Lemma 4.1 *Let $\{t_m\}_{m=1}^{\infty}$ be a positive, increasing real sequence such that $t_m \longrightarrow \infty$ as $m \longrightarrow \infty$. For every $m \geq 1$, let M_m be a symmetric, nonnegative definite $p \times p$ matrix with ordered eigenvalues $0 \leq \lambda_{m,p} \leq \cdots \leq \lambda_{m,1}$ and corresponding orthonormal eigenvectors $\eta_{m,i}$, $i = 1, \ldots, p$. Assume that*

$$\lim_{m \to \infty} \frac{M_m}{t_m} = \sum_{i=1}^{k} c_i \cdot \beta_i \beta_i^{\top} \tag{4.2}$$

for some $0 < c_k \leq \cdots \leq c_1$ and orthonormal collection of vectors $\beta_i \in \mathbb{R}^p$, $i = 1, \ldots, k$, $k \leq p$. Furthermore, suppose that there are $g \leq k$ distinct values among $\{c_1, \ldots, c_k\}$, denoted by $0 < \widetilde{c}_g < \cdots < \widetilde{c}_1$. For each $j = 1, \ldots, g$, define an index set $\widetilde{C}_j = \{i : c_i = \widetilde{c}_j\}$, write $\widetilde{B}_j$ and $\widetilde{N}_{m,j}$ to denote $p \times |\widetilde{C}_j|$ matrices with columns formed by vectors $\beta_i, i \in \widetilde{C}_j$ and $\eta_{m,i}, i \in \widetilde{C}_j$, respectively; denote projection matrices for $\widetilde{B}_j$ and $\widetilde{N}_{m,j}$ as $P_{\widetilde{B}_j}$ and $P_{\widetilde{N}_{m,j}}$, respectively.
(a) Then, for each $j = 1, \ldots, g$,

$$\lim_{m \to \infty} P_{\widetilde{N}_{m,j}} = P_{\widetilde{B}_j}.$$

There exists a sequence of $k \times k$ orthogonal matrices

$$
Q_m^* = \begin{bmatrix}
\widetilde{N}_{m,1}^\top \widetilde{B}_1 & 0_{|\widetilde{B}_1| \times |\widetilde{B}_2|} & 0_{|\widetilde{B}_1| \times |\widetilde{B}_3|} & \cdots & 0_{|\widetilde{B}_1| \times |\widetilde{B}_g|} \\
0_{|\widetilde{B}_2| \times |\widetilde{B}_1|} & \widetilde{N}_{m,2}^\top \widetilde{B}_2 & 0_{|\widetilde{B}_2| \times |\widetilde{B}_3|} & \cdots & 0_{|\widetilde{B}_2| \times |\widetilde{B}_g|} \\
\vdots & \vdots & \ddots & \cdots & \vdots \\
0_{|\widetilde{B}_g| \times |\widetilde{B}_1|} & 0_{|\widetilde{B}_g| \times |\widetilde{B}_2|} & 0_{|\widetilde{B}_g| \times |\widetilde{B}_3|} & \cdots & \widetilde{N}_{m,g}^\top \widetilde{B}_g
\end{bmatrix}
$$

such that, as $m \longrightarrow \infty$,

$$
\begin{bmatrix} \widetilde{N}_{m,1} \cdots \widetilde{N}_{m,g} \end{bmatrix} Q_m^* \longrightarrow \begin{bmatrix} \widetilde{B}_1 \cdots \widetilde{B}_g \end{bmatrix}.
$$

(b) Let $\mathcal{N}_m$ and $\mathcal{B}$ denote subspaces of $\mathbb{R}^p$ spanned by $\eta_{m,i}$ and β_i vectors, $i = 1, \ldots, k$, respectively. Then, as $m \longrightarrow \infty$,

$$
P_{\mathcal{N}_m} = \sum_{i=1}^{k} \eta_{m,i} \eta_{m,i}^\top \longrightarrow \sum_{i=1}^{k} \beta_i \beta_i^\top = P_{\mathcal{B}}.
$$

There exists a sequence Q_m of $k \times k$ orthogonal matrices such that, as $m \longrightarrow \infty$,

$$
\begin{bmatrix} \eta_{m,1} \cdots \eta_{m,k} \end{bmatrix} Q_m \longrightarrow \begin{bmatrix} \beta_1 \cdots \beta_k \end{bmatrix}.
$$

(c) For $i = 1, \ldots, p$,

$$
\lim_{m \to \infty} \frac{\lambda_{m,i}}{t_m} = \begin{cases} c_i & \text{if } i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases}
$$

Remark. We make a few comments on the above result before beginning its proof. Under the matrix convergence from (4.2), Lemma 4.1 states that the ordered eigenvalues of M_m (scaled by t_m), as well as the corresponding eigenvectors, converge to those of the matrix $\sum_{i=1}^{k} c_i \cdot \beta_i \beta_i^\top$. While the first k ordered eigenvalues of M_m converge directly to the corresponding ordered eigenvalues of the limiting matrix $\sum_{i=1}^{k} c_i \cdot \beta_i \beta_i^\top$ by part (c), the convergence of the eigenvectors of M_m is *not* direct; that is, we cannot generally say that $\eta_{m,i} \to \beta_i$ $(i = 1, \ldots, k)$. But instead, by Lemma 4.1 (b), the first k orthonormal eigenvectors of M_m converge to those of the limiting matrix, namely $\beta_1, \ldots, \beta_k$, up to an orthogonal transformation. Consequently, the space $\mathcal{N}_m$ spanned by the k vectors $\eta_{m,i}$ (corresponding to the k largest eigenvalues of M_m) converges as

$m \to \infty$ to the space $\mathcal{B}$ spanned by the eigenvectors β_i, $i = 1, \ldots, k$ of the limiting matrix (4.2). As an implication of this, direct componentwise convergence of the first k eigenvectors of M_m of the form $\eta_{m,i} \to \beta_i$ $(i = 1, \ldots, k)$ can be achieved in fact if the first k eigenvalues of (4.2) are all distinct $c_k < \cdots < c_1$, or equivalently if $g = k$ and $|\widetilde{C}_j| = 1$ for all $j = 1, \ldots, g$. In part (a), we also find that if there are g distinct eigenvalues among $\{c_i\}_{i=1}^k$ in (4.2) then, up to an orthogonal transformation, g different *collections* of eigenvectors of M_m must converge to those *collections* of β_i vectors in (4.2) which share a common c_i eigenvalue. Because the limiting matrix in (4.2) has rank k, part (c) of Lemma 4.1 reflects that the smallest $p - k$ eigenvalues of the matrix M_m must converge to zero, ensuring the rank of M_m properly converges to k.

Proof. We separately treat the proofs of parts (a)-(c) for clarity.

Proof of (c).

We begin by establishing part (c) first. Note that the symmetric matrix M_m/t_m has ordered eigenvalues $\lambda_{m,i}$, $i = 1, \ldots, p$, which correspond to the orthonormal eigenvectors $\eta_{m,i}$. Also, the limiting matrix $\sum_{i=1}^k c_i \beta_i \beta_i^\top$ from (4.2) has ordered positive eigenvalues $c_1 > \cdots > c_k > 0$ along with an eigenvalue 0 with multiplicity $p - k$ (i.e., exactly $p - k$ times does 0 appear as an eigenvalue), corresponding to $p - k$ additional orthonormal eigenvectors, say $\beta_{k+1}, \ldots, \beta_p$, that span the orthogonal complement of $\text{span}\{\beta_i\}_{i=1}^k$, i.e., $\sum_{i=1}^k c_i \beta_i \beta_i^\top + 0 \sum_{i=k+1}^p \beta_i \beta_i^\top$. By Rouché's theorem [Stewart and Sun (1990), p. 167], which states that eigenvalues are continuous under matrix convergence, it holds that the ordered eigenvalues of M_m/t_m must converge to those of the limiting matrix (4.2) so that

$$\lim_{m \to \infty} \frac{\lambda_{m,i}}{t_m} = \begin{cases} c_i & \text{if } i = 1, \ldots, k, \\ 0 & \text{otherwise} \end{cases}$$

follows for $i = 1, \ldots, p$. In particular, the convergence of the first k ordered eigenvalues of M_m/t_m implies that

$$\lim_{m \to \infty} \frac{\lambda_{m,i}}{t_m} = \widetilde{c}_i, \qquad i \in \widetilde{C}_j, \ j = 1, \ldots, g. \tag{4.3}$$

Proof of (a).

For each $j = 1, \ldots, g$, note that $P_{\widetilde{N}_{m,j}} = \sum_{i \in \widetilde{C}_j} \eta_{m,i} \eta_{m,i}^\top$ and $P_{\widetilde{B}_j} = \sum_{i \in \widetilde{C}_j} \beta_i \beta_i^\top$ by

the orthonormal property of the vectors $\eta_{m,i}$ and β_i, respectively. It also holds that $M_m = \sum_{i=1}^{p} \lambda_{m,i} \cdot \eta_{m,i} \eta_{m,i}^\top$ and $\sum_{i=1}^{k} c_i \cdot \beta_i \beta_i^\top = \sum_{j=1}^{g} \widetilde{c}_j P_{\widetilde{B}_j}$. The matrix M_m can be rewritten as

$$M_m = t_m \sum_{i=1}^{k} c_i \beta_i \beta_i^\top + R_m, \tag{4.4}$$

where the remainder matrix R_m is defined by subtraction, i.e. $R_m = M_m - t_m \sum_{i=1}^{k} c_i \beta_i \beta_i^\top$. Let $t_m \cdot r_{m,0}^+$ denote the sum of the absolute entries of R_m; that is, if the $p \times p$ matrix R_m has entries denoted by $R_{m,ij}$, $i = 1, \ldots, p, j = 1, \ldots, p$, then $t_m \cdot r_{m,0}^+ = \sum_{i=1}^{p} \sum_{j=1}^{p} |R_{m,ij}|$. Because any component of $\eta_{m,i}$ is less than or equal to 1 in absolute value by $\|\eta_{m,i}\| = 1$ for $i = 1, \ldots, k$, we may bound

$$|\eta_{m,i}^\top R_m \eta_{m,i}| \leq t_m \cdot r_{m,0}^+, \quad i = 1, \ldots, k, \quad \text{where} \quad \lim_{m \to \infty} r_{m,0}^+ = 0 \tag{4.5}$$

follows from $\lim_{m \to \infty} R_m/t_m = 0_{p \times p}$.

To establish the claim in part (a), it suffices to show that: for any $j = 1, \ldots, g$,

$$\lim_{m \to \infty} \|(\mathrm{I}_p - P_{\widetilde{B}_j}) \eta_{m,i}\| = 0, \qquad i \in \widetilde{C}_j, \tag{4.6}$$

denoting the $p \times p$ identity matrix as I_p above. From (4.6), it follows that $\lim_{m \to \infty}(\mathrm{I}_p - P_{\widetilde{B}_j}) \eta_{m,i} = 0_p$ for any $i \in \widetilde{C}_j$ and fixed $j = 1, \ldots, g$ implying that, when m is large, we have $\eta_{m,i} \approx P_{\widetilde{B}_j} \eta_{m,i}$ for all $i \in \widetilde{C}_j$ for a given $j = 1, \ldots, g$.

Because components of $\eta_{m,i}$ are bounded by 1 from $\|\eta_{m,i}\| = 1$, we find that, if (4.6) holds for a given $j = 1, \ldots, g$, then

$$\lim_{m \to \infty} (\mathrm{I}_p - P_{\widetilde{B}_j}) P_{\widetilde{N}_{m,j}} = \lim_{m \to \infty} \sum_{i \in \widetilde{C}_j} (\mathrm{I}_p - P_{\widetilde{B}_j}) \eta_{m,i} \eta_{m,i}^\top = 0_{p \times p} \tag{4.7}$$

holds as well. Because $\widetilde{N}_{m,i}^\top \widetilde{N}_{m,i} = \mathrm{I}_{|\widetilde{C}_j|}$ (the $|\widetilde{C}_j| \times |\widetilde{C}_j|$ identity matrix), $P_{\widetilde{B}_j} = \widetilde{B}_j \widetilde{B}_j^\top$ and $P_{\widetilde{N}_{m,j}} = \widetilde{N}_{m,i} \widetilde{N}_{m,i}^\top$, we find (4.7) is in turn equivalent to

$$0_{|\widetilde{C}_j| \times |\widetilde{C}_j|} = \lim_{m \to \infty} \widetilde{N}_{m,j}^\top (\mathrm{I}_{|\widetilde{C}_j|} - P_{\widetilde{B}_j}) P_{\widetilde{N}_{m,j}} \widetilde{N}_{m,j} = \lim_{m \to \infty} (\mathrm{I}_{|\widetilde{C}_j|} - A_m A_m^\top), \tag{4.8}$$

where $A_m = \widetilde{N}_{m,j}^\top \widetilde{B}_j$ for a given $j = 1, \ldots, g$. By Rouché's theorem again, all $|\widetilde{C}_j|$ eigenvalues of $A_m A_m^\top$ must converge to 1 as $m \to \infty$ so that

$$\begin{aligned} \sup_m \|A_m\|_2^2 &= \sup_m (\text{largest eigenvalue of } A_m A_m^\top) &\leq C \\ \sup_m \|A_m^{-1}\|_2^2 &= \sup_m 1/(\text{smallest eigenvalue of } A_m A_m^\top) &\leq C \end{aligned} \tag{4.9}$$

for some real $0 < C < \infty$. In addition, since $\det(A_m) = \det(A_m^\top)$, we have

$$1 = \lim_{m \to \infty} \det(A_m A_m^\top) = \lim_{m \to \infty} \det(A_m) \cdot \det(A_m^\top) = \lim_{m \to \infty} \left[\det(A_m) \right]^2,$$

which shows that $\lim_{m \to \infty} \det(A_m) = 1$ so that A_m is nonsingular for large m and A_m^{-1} exists eventually. Then, applying Lemma 7.4 with (4.8) and (4.9), we find

$$
\begin{aligned}
\lim_{m \to \infty} \| \mathrm{I}_{|\widetilde{C}_j|} - A_m^\top A_m \|_{\mathbb{F}} &= \lim_{m \to \infty} \| A_m^{-1} A_m (\mathrm{I}_{|\widetilde{C}_j|} - A_m^\top A_m) \|_{\mathbb{F}} \\
&= \lim_{m \to \infty} \| A_m^{-1} (\mathrm{I}_{|\widetilde{C}_j|} - A_m A_m^\top) A_m \|_{\mathbb{F}} \\
&\leq \lim_{m \to \infty} \| A_m^{-1} \|_{\mathbb{F}} \cdot \| (\mathrm{I}_{|\widetilde{C}_j|} - A_m A_m^\top) \|_{\mathbb{F}} \cdot \| A_m \|_{\mathbb{F}} \\
&\leq \lim_{m \to \infty} |\widetilde{C}_j| \| A_m^{-1} \|_2 \cdot \| (\mathrm{I}_{|\widetilde{C}_j|} - A_m A_m^\top) \|_{\mathbb{F}} \cdot \| A_m \|_2 \\
&\leq |\widetilde{C}_j| \cdot C \cdot \lim_{m \to \infty} \| (\mathrm{I}_{|\widetilde{C}_j|} - A_m A_m^\top) \|_{\mathbb{F}} \\
&= 0,
\end{aligned}
$$

so that $\lim_{m \to \infty} (\mathrm{I}_{|\widetilde{C}_j|} - A_m^\top A_m) = 0_{|\widetilde{C}_j| \times |\widetilde{C}_j|}$ and hence

$$
\begin{aligned}
0_{p \times p} &= \lim_{m \to \infty} \widetilde{B}_j (\mathrm{I}_p - A_m^\top A_m) \widetilde{B}_j^\top \\
&= \lim_{m \to \infty} P_{\widetilde{B}_j} - P_{\widetilde{B}_j} P_{\widetilde{N}_{m,j}} P_{\widetilde{B}_j} \\
&= \lim_{m \to \infty} \left[(\mathrm{I}_p - P_{\widetilde{N}_{m,j}}) P_{\widetilde{B}_j} \right]^\top \left[(\mathrm{I}_p - P_{\widetilde{N}_{m,j}}) P_{\widetilde{B}_j} \right],
\end{aligned}
$$

because $\mathrm{I}_p - P_{\widetilde{N}_{m,j}}$ and $P_{\widetilde{B}_j}$ are symmetric, idempotent matrices. This last limit shows that, if (4.6) holds for a given $j = 1, \ldots, g$, then

$$\lim_{m \to \infty} (\mathrm{I}_p - P_{\widetilde{N}_{m,j}}) P_{\widetilde{B}_j} = 0_{p \times p}, \tag{4.10}$$

holds in addition to the counterpart result in (4.7). From (4.7) and (4.10), we find the following limit of the quadratic form

$$
\begin{aligned}
\lim_{m \to \infty} (P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}})^\top (P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}}) &= \lim_{m \to \infty} \left[(\mathrm{I}_p - P_{\widetilde{N}_{m,j}}) P_{\widetilde{B}_j} + (\mathrm{I}_p - P_{\widetilde{B}_j}) P_{\widetilde{N}_{m,j}} \right] \\
&= 0_{p \times p}
\end{aligned}
$$

so that

$$\lim_{m \to \infty} (P_{\widetilde{N}_{m,j}} - P_{\widetilde{B}_j}) = 0_{p \times p} \tag{4.11}$$

is valid whenever (4.6) holds for a given $j = 1, \ldots, g$. This shows the first claim in part (a) of Lemma 4.1. The second claim in Lemma 4.1(a) then follows using

$$\left[\widetilde{N}_{m,1} \cdots \widetilde{N}_{m,g}\right] Q_m^* = \left[P_{\widetilde{N}_{m,1}} \widetilde{B}_1 \cdots P_{\widetilde{N}_{m,g}} \widetilde{B}_g\right],$$

so that the limiting matrix

$$
\begin{aligned}
\lim_{m \to \infty} \left[\widetilde{N}_{m,1} \cdots \widetilde{N}_{m,g}\right] Q_m^* &= \lim_{m \to \infty} \left[P_{\widetilde{N}_{m,1}} \widetilde{B}_1 \cdots P_{\widetilde{N}_{m,g}} \widetilde{B}_g\right] \\
&= \left[P_{\widetilde{B}_1} \widetilde{B}_1 \cdots P_{\widetilde{B}_g} \widetilde{B}_g\right] \\
&= \left[\widetilde{B}_1 \cdots \widetilde{B}_g\right],
\end{aligned}
$$

because $P_{\widetilde{B}_j} \widetilde{B}_j = \widetilde{B}_j$ for $j = 1, \ldots, g$.

Proof of (4.6). We use an induction argument on j to establish (4.6). We first show that (4.6) holds for $j = 1$ and any $i = 1, \ldots, |\widetilde{C}_1|$ (that is, $i \in \widetilde{C}_1$).

Using the orthonormal property of the eigenvectors of M_m, (4.4) and (4.5), we find for $i \in \widetilde{C}_1$:

$$
\begin{aligned}
0 &\leq \lambda_{m,i} \\
&= \eta_{m,i}^\top M_m \eta_{m,i} \\
&\leq t_m \sum_{\ell=1}^{g} \widetilde{c}_\ell \cdot \eta_{m,i}^\top P_{\widetilde{B}_\ell} \eta_{m,i} + |\eta_{m,i}^\top R_m \eta_{m,i}| \\
&\leq t_m \sum_{\ell=1}^{g} \widetilde{c}_\ell \cdot \|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 + t_m \cdot r_{m,0}^+ \qquad (4.12) \\
&\leq t_m \widetilde{c}_1 \sum_{\ell=1}^{g} \|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 + t_m \cdot r_{m,0}^+ \\
&= t_m \widetilde{c}_1 \eta_{m,i}^\top P_{\mathcal{B}} \eta_{m,i} + t_m \cdot r_{m,0}^+ \\
&\leq t_m \widetilde{c}_1 + t_m \cdot r_{m,0}^+,
\end{aligned}
$$

where the last inequality follows from:

$$\eta_{m,i}^\top P_{\mathcal{B}} \eta_{m,i} \leq \eta_{m,i}^\top P_{\mathcal{B}} \eta_{m,i} + \eta_{m,i}^\top (I_p - P_{\mathcal{B}}) \eta_{m,i} = \eta_{m,i}^\top \eta_{m,i} = 1.$$

Because $\widetilde{c}_1 = c_i$ for $i \in \widetilde{C}_1$ and

$$\lim_{m \to \infty} \frac{t_m \cdot r_{m,0}^+}{t_m} = 0, \qquad \lim_{m \to \infty} \frac{\lambda_{m,i}}{t_m} = c_i = \widetilde{c}_1$$

by (4.3) and (4.5), we may conclude that

$$\lim_{m \to \infty} \sum_{\ell=1}^{g} \|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 = 1, \tag{4.13}$$

from the previous inequalities in (4.12). We aim now to show that for $i \in \widetilde{C}_1$

$$\lim_{m \to \infty} \|P_{\widetilde{B}_1} \eta_{m,i}\|^2 = 1 \tag{4.14}$$

follows from (4.13). If $g = 1$, this is clearly true. For the case that $g > 1$, it suffices to establish for a fixed $i \in \widetilde{C}_1$,

$$\lim_{m \to \infty} \|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 = 0, \quad \ell = 2, \ldots, g, \tag{4.15}$$

or equivalently, (because the sequence $\|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 \leq \|\eta_{m,i}\|^2 = 1$ is bounded by the Cauchy-Schwartz inequality) that any convergent subsequence $a_{\ell,n} = \|P_{\widetilde{B}_\ell} \eta_{m_n,1}\|^2$ of $\|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2$ converges to 0, for any $\ell = 2, \ldots, g$. Suppose (4.15) does not hold and, for some $\ell^* \in \{2, \ldots, g\}$, there exists a subsequence $a_{\ell^*,n}$ such that $a_{\ell^*,n} \longrightarrow c \neq 0$ as $n \longrightarrow \infty$. Note that $c \in (0, 1]$ from $0 \leq a_{\ell^*,n} \leq 1$ and that by (4.13): $i \in \widetilde{C}_1$,

$$\lim_{n \to \infty} \sum_{\substack{\ell=1 \\ \ell \neq \ell^*}}^{g} \|P_{\widetilde{B}_\ell} \eta_{m_n,i}\|^2 = \lim_{n \to \infty} \left(\sum_{\ell=1}^{g} \|P_{\widetilde{B}_\ell} \eta_{m_n,i}\|^2 - a_{\ell^*,n} \right) = 1 - c.$$

Using (4.12), we can produce bounds

$$\frac{\lambda_{m_n,i}}{t_{m_n}} \leq \widetilde{c}_{\ell^*} a_{\ell^*,n} + \widetilde{c}_1 \sum_{\substack{\ell=1 \\ \ell \neq \ell^*}}^{g} \|P_{\widetilde{B}_\ell} \eta_{m_n,1}\|^2 + \frac{t_{m_n} \cdot r_{m_n,0}^+}{t_{m_n}} \leq \widetilde{c}_1 + r_{m_n,0}^+.$$

Taking limits as $n \to \infty$, we then find $\widetilde{c}_1 \leq c \cdot \widetilde{c}_{\ell^*} + (1 - c) \cdot \widetilde{c}_1 \leq \widetilde{c}_1$, implying $\widetilde{c}_1 = \widetilde{c}_{\ell^*}$. This is a contradiction because $\widetilde{c}_{\ell^*} < \widetilde{c}_1$ for $\ell^* \neq 1$ by assumption. Hence, we have now established that (4.14) holds or equivalently that (4.6) holds with $j = 1$.

If $g = 1$, we have immediately that (4.6) holds. For the case $g > 1$, we now use a strong induction step to show that, if (4.6) holds for any $j \leq j^*$ for some fixed j^* where $1 \leq j^* \leq g - 1$ (the induction assumption), then (4.6) also holds for $j = j^* + 1$. For any j with $j \leq j^*$, if (4.6) holds then $\lim_{m \to \infty} (P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}}) = 0_{p \times p}$ follows from (4.11).

Let $r_{m,j}^+$ denote the sum of the absolute entries of $P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}}$, $j = 1, \ldots, g$. Again, the components of the eigenvectors $\eta_{m,i}$ are bounded by 1, so that we may majorize

$$|\eta_{m,i}^\top (P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}})\eta_{m,i}| \leq r_{m,j}^+, \quad i \in \widetilde{C}_j, \quad \text{where } \lim_{m \to \infty} r_{m,j}^+ = 0, \quad j \leq j^*, \quad (4.16)$$

noting that $\lim_{m \to \infty} r_{m,j}^+ = 0$ follows for $j \leq j^*$ from $\lim_{m \to \infty}(P_{\widetilde{B}_j} - P_{\widetilde{N}_{m,j}}) = 0_{p \times p}$ by the induction assumption that (4.6) and (4.11) hold for $j \leq j^*$.

Now let $j = j^* + 1$ and $i \in \widetilde{C}_{j^*+1}$, or equivalently, $1 \leq i - \sum_{\ell=1}^{j^*} |\widetilde{C}_\ell| \leq |\widetilde{C}_{j^*+1}|$. By steps analogous to (4.12) using (4.16), $\eta_{m,i}^\top \eta_{m,i} = 1$ and $P_{\widetilde{N}_{m,j}}\eta_{m,i} = 0_p$ for $j \leq j^*$ and $i \in \widetilde{C}_{j^*+1}$, we can argue for $i \in \widetilde{C}_{j^*+1}$,

$$
\begin{aligned}
0 \;\leq\;& \lambda_{m,i} \\
=\;& \eta_{m,i}^\top M_{m,i} \eta_{m,i} \hspace{6cm} (4.17) \\
\leq\;& t_m \sum_{\ell=1}^{j^*} \widetilde{c}_\ell \cdot \eta_{m,i}^\top \big[(P_{\widetilde{B}_\ell} - P_{\widetilde{N}_{m,\ell}}) + P_{\widetilde{N}_{m,\ell}}\big]\eta_{m,i} \\
& + t_m \sum_{\ell=j^*+1}^{g} \widetilde{c}_\ell \cdot \eta_{m,i}^\top P_{\widetilde{B}_\ell}\eta_{m,i} + |\eta_{m,i}^\top R_m \eta_{m,i}| \hspace{1.5cm} (4.18) \\
\leq\;& t_m \sum_{\ell=j^*+1}^{g} \widetilde{c}_\ell \cdot \|P_{\widetilde{B}_\ell}\eta_{m,i}\|^2 + t_m \cdot \sum_{\ell=0}^{j^*} \widetilde{c}_\ell \cdot r_{m,\ell}^+ \quad (\widetilde{c}_0 = 1) \\
\leq\;& t_m \widetilde{c}_{j^*+1} \sum_{\ell=j^*+1}^{g} \|P_{\widetilde{B}_\ell}\eta_{m,i}\|^2 + t_m \sum_{\ell=0}^{j^*} \widetilde{c}_\ell \cdot r_{m,\ell}^+ \\
\leq\;& t_m \widetilde{c}_{j^*+1} + t_m \sum_{\ell=0}^{j^*} \widetilde{c}_\ell \cdot r_{m,\ell}^+.
\end{aligned}
$$

Above we used $\widetilde{c}_{j^*+1} > \widetilde{c}_j > 0$ for $j > j^*$. From (4.3), (4.5), (4.16) and $\widetilde{c}_{j^*+1} = c_i$ for $i \in \widetilde{C}_{j^*+1}$, it follows for $i \in \widetilde{C}_{j^*+1}$ that

$$\lim_{m \to \infty} \sum_{\ell=0}^{j^*} \widetilde{c}_\ell \cdot r_{m,\ell}^+ = 0, \quad \lim_{m \to \infty} \frac{\lambda_{m,i}}{t_m} = c_i = \widetilde{c}_{j^*+1},$$

and we find in (4.18) that

$$\lim_{m \to \infty} \sum_{\ell=j^*+1}^{g} \|P_{\widetilde{B}_\ell}\eta_{m,i}\|^2 = 1, \quad i \in \widetilde{C}_{j^*+1}. \quad (4.19)$$

To finish the induction argument, we wish to show that for $i \in \widetilde{C}_{j^*+1}$

$$\lim_{m \to \infty} \|P_{\widetilde{B}_{j^*+1}} \eta_{m,i}\|^2 = 1$$

now follows from (4.19). If $g = j^* + 1$, this is immediate and, for the case that $g > j^* + 1$, we can establish for $i \in \widetilde{C}_{j^*+1}$

$$\lim_{m \to \infty} \|P_{\widetilde{B}_\ell} \eta_{m,i}\|^2 = 0, \quad \ell = j^* + 2, \ldots, g,$$

with arguments completely analogous to (4.15) involving subsequences. (If for some $\ell^* \in \{j^* + 2, \ldots, g\}$ and $i \in \widetilde{C}_{j^*+1}$, there exists a non-null subsequence $b_{m_n,\ell^*} = \|P_{\widetilde{B}_{\ell^*}} \eta_{m_n,i}\|^2$ with $\lim_{n \to \infty} b_{m_n,\ell^*} = c \in (0,1]$, then we may deduce from (4.18) and (4.19) by taking limits as $n \to \infty$ that $\widetilde{c}_{j^*+1} \leq c \cdot \widetilde{c}_{\ell^*} + (1-c) \cdot \widetilde{c}_{j^*+1} \leq \widetilde{c}_{j^*+1}$, a contradiction since $\widetilde{c}_{\ell^*} < \widetilde{c}_{j^*+1}$.) The proof of (4.6) by induction is now complete.

Proof of (b). The convergence in part (b) follows immediately from the result in part (a) and

$$P_{\mathcal{N}_m} = \sum_{j=1}^{g} P_{\widetilde{N}_{m,j}}, \qquad P_{\mathcal{B}} = \sum_{j=1}^{g} P_{\widetilde{B}_j}$$

by orthogonality. $\square$

Data contamination

Suppose the model assumption

$$Y = f(B^\top X, \varepsilon)$$

with $\mu = \mathrm{E}(X)$ and $\Sigma = \mathrm{Cov}(X)$ can be justified and $(X, Y)^n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ is a sample of size n of (X, Y) with $y_i \neq y_j$ for all $i \neq j$. When the dimension $1 \leq \mathcal{K} < p$ of the reduction subspace and the covariance structure Σ, μ are known, a dimension reduction procedure of the Sir-type applied to $(X, Y)^n$ yields e.d.r. directions $\widehat{\beta}_i = \Sigma^{-1/2} \widehat{\eta}_i$, $i = 1, \ldots, \mathcal{K}$ derived from the orthonormal eigenvectors $\widehat{\eta}_i$, $i = 1, \ldots, \mathcal{K}$ of

$$\widehat{V} = \Sigma^{-1/2} \sum_{h=1}^{H} \widehat{p}_h (\overline{x}_h - \mu)(\overline{x}_h - \mu)^\top \Sigma^{-1/2} \tag{4.20}$$

corresponding to the $\mathcal{K}$ largest, ordered eigenvalues $\widehat{\lambda}_{\mathcal{K}} \leq \cdots \leq \widehat{\lambda}_1$ of (4.20). The estimated reduction space $\widehat{\mathcal{B}}$ is spanned by the vectors $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$.

Recall that, in the SIR procedure, the sample observations of $\{(x_i, y_i)\}_{i=1}^n$ are ordered by the $y-$responses ($y_{(i)} \leq y_{(i+1)}$) and grouped into slices $\mathbb{I}_i$, $i = 1, \ldots, H$ as determined by the rankings of the response variable:

$$y_{(1)}, \ldots, y_{(n_1)} \in \mathbb{I}_1; \quad y_{(n_1+1)}, \ldots, y_{(n_1+n_2)} \in \mathbb{I}_2; \quad \cdots \quad y_{(n_1+\cdots+n_{H-1}+1)}, \ldots, y_{(n)} \in \mathbb{I}_H.$$

We shall exploit the slices in the contamination scheme as follows, noting $\mathcal{K} \leq H$.

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 scaling factor t_m, we create a contaminated sample, denoted by $(X, Y)_m^{n,k}$, by replacing k observations in the uncontaminated data $(X, Y)^n$. Denote the index of the first observation in each slice $\mathbb{I}_h$ as $1_h = 1 + \sum_{i=1}^{h-1} n_i$, $h = 1, \ldots, H$. WLoG 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

$$(\widetilde{x}_{1_h, m}, \widetilde{y}_{1_h}), \quad h = 1, \ldots, k,$$

where the y-values remain unchanged, $\widetilde{y}_{1_h} = y_{1_h}$, $h = 1, \ldots, k$ and thus observations assigned to a slice $\mathbb{I}_h, h = 1, \ldots, H$ stay the same. The contamination of the x-values is assumed to be of the following structure:

$$\widetilde{x}_{1_h, m} = t_m \widetilde{\beta}_h + \widetilde{v}_h, \quad \widetilde{\beta}_h, \widetilde{v}_h \in \mathbb{R}^p, \quad h = 1, \ldots, k; \tag{4.21}$$

with

$$\widetilde{\beta}_h^\top \Sigma^{-1} \widetilde{\beta}_j = \delta_{hj}, \quad h, j = 1, \ldots, k, \tag{4.22}$$

where δ_{hj} denotes the Kronecker delta function above (e.g., $\delta_{hj} = 1$ if $j = h$ and otherwise 0). Again, under this $k-$observation replacement scheme, we obtain the contaminated sample as $(X, Y)_m^{n,k}$.

Computing a version of (4.20) based on $(X, Y)_m^{n,k}$ yields a contaminated estimate $\widehat{V}_m$ of $\mathrm{Cov}(\mathrm{E}(\Sigma^{-1/2} X | Y))$ with ordered eigenvalues $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ and corresponding orthonormal eigenvectors $\widehat{\eta}_{m,i}$, $i = 1, \ldots, p$ of $\widehat{V}_m$. An application of the SIR

method to $(X, Y)_m^{n,k}$ results in an estimate $\widehat{\mathcal{B}}_{k,m}$ of the dimension reduction subspace spanned by an appropriate choice of vectors $\widehat{\beta}_{m,i} = \Sigma^{-1/2}\widehat{\eta}_{m,i}$, $i = 1, \ldots, \mathcal{K}$.

With the established Lemma 4.1, we can now study the effect of this type of contamination on the eigenvectors of $\widehat{V}_m$ in determining estimates of e.d.r. directions. The following theorem quantifies the result.

Theorem 4.1 *Under the k-slice contamination scheme of Section 4.3.1, $1 \le k \le \min\{p, H\}$, and with known $\mathrm{E}(X) = \mu$ and $\mathrm{Cov}(X) = \Sigma$, suppose that a SIR-dimension reduction procedure is applied to the contaminated sample $(X, Y)_m^{n,k}$ with resulting ordered eigenvalues $0 \le \widehat{\lambda}_{m,p} \le \cdots \le \widehat{\lambda}_{m,1}$ and corresponding orthonormal eigenvectors $\widehat{\eta}_{m,i}$, $i = 1, \ldots, p$ of an estimated covariance matrix $\widehat{V}_m$.*

(a) Let $\widehat{\mathcal{N}}_m$ and $\Sigma^{-1/2}\widetilde{\mathcal{B}}$ denote subspaces of $\mathbb{R}^p$ spanned by $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ and $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^k$ vectors, respectively. Then, as $m \longrightarrow \infty$,

$$P_{\widehat{\mathcal{N}}_m} = \sum_{h=1}^k \widehat{\eta}_{m,h}\widehat{\eta}_{m,h}^\top \longrightarrow \sum_{h=1}^k (\Sigma^{-1/2}\widetilde{\beta}_{m,h})(\Sigma^{-1/2}\widetilde{\beta}_{m,h})^\top = P_{\Sigma^{-1/2}\widetilde{\mathcal{B}}}.$$

(b) If the number of observations in the first k slices, namely $\{n_h\}_{h=1}^k$, are ordered $n_{(1)} \le \cdots \le n_{(k)}$, then

$$\lim_{m \to \infty} \frac{\widehat{\lambda}_{m,h}}{t_m^2} = \begin{cases} \dfrac{1}{n_{(k-h+1)}n} & \text{if } h = 1, \ldots, k, \\[2mm] 0 & \text{otherwise.} \end{cases}$$

(c) For $h \in \{1, \ldots, k\}$, let $\widehat{\mathcal{N}}_{m,h}$ and $\Sigma^{-1/2}\widetilde{\mathcal{B}}_h$ denote subspaces of $\mathbb{R}^p$ spanned by $\{\widehat{\eta}_{m,j} : 1 \le j \le k, \ n_{(k-j+1)} = n_{(k-h+1)}\}$ and $\{\Sigma^{-1/2}\widetilde{\beta}_j : 1 \le j \le k, \ n_j = n_{(k-h+1)}\}$, respectively. As $m \longrightarrow \infty$,

$$P_{\widehat{\mathcal{N}}_{m,h}} = \sum_{\substack{1 \le j \le k, \\ n_{(k-j+1)} = n_{(k-h+1)}}} \widehat{\eta}_{m,h}\widehat{\eta}_{m,h}^\top \longrightarrow \sum_{\substack{1 \le j \le k, \\ n_j = n_{(k-h+1)}}} (\Sigma^{-1/2}\widetilde{\beta}_{m,h})(\Sigma^{-1/2}\widetilde{\beta}_{m,h})^\top = P_{\Sigma^{-1/2}\widetilde{\mathcal{B}}_h}.$$

(d) There exists a sequence Q_m of orthogonal $k \times k$ matrices such that

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

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

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

Remark. We start again with a few comments on the above results before proceeding with the proof. Under the k-slice contamination scheme of Section 4.3.1, a contaminated covariance matrix $\widehat{V}_m$ results in place of (4.20) by using the corrupted data $(X, Y)_m^{n,k}$. This matrix has orthonormal eigenvectors $\widehat{\eta}_{m,p}, \ldots, \widehat{\eta}_{m,1}$ associated with the ordered eigenvalues $\widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$. In particular, the k largest eigenvalues of $\widehat{V}_m$ correspond to the eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$. Part (a) of Theorem 4.1 states that $\mathbb{R}^p$-space $\widehat{\mathcal{N}}_m$ spanned by the contaminated eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ converges to the $\mathbb{R}^p$-space $\widetilde{\mathcal{B}}$ spanned by $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^k$ based on the k directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^k$ satisfying (4.22) (i.e., the respective projection matrices converge). In matrix form, Theorem 4.1(d) states that, up to an orthogonal transformation, the contaminated eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ converge to $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^k$. Consequently, the space $\Sigma^{-1/2}\widehat{\mathcal{N}}_m$ spanned by contaminated first k e.d.r. directions $\{\Sigma^{-1/2}\widehat{\eta}_{m,h}\}_{h=1}^k$ (i.e., transforming the eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ associated with the k largest eigenvalues of $\widehat{V}_m$, where we note that k may not necessarily equal $\mathcal{K}$) converges to the space $\Sigma^{-1}\widetilde{\mathcal{B}}$ spanned by $\{\Sigma^{-1}\widetilde{\beta}_h\}_{h=1}^k$ in Theorem 4.1(e). This result is crucial because it reveals how the k-slice contamination scheme of Section 4.3.1 can be used to control vectors spanning the e.d.r. subspace. In part (b) of Theorem 4.1, we find that, after contaminating k slices, the k largest eigenvalues of the contaminated matrix $\widehat{V}_m$ "explode" and grow infinitely large at a rate faster than the remaining $p - k$ eigenvalues of $\widehat{V}_m$. (In the case where the covariance structure Σ of X is known but the dimension $\mathcal{K}$ of the true e.d.r. subspace is *unknown*, this result implies that we can carefully control the dimension of the estimated e.d.r. subspace, as well as the space itself under contamination by determining how many eigenvalues are *large*; see Section 4.4.2.) Finally, Theorem 4.1(c) states that, because the eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ associated with the k largest eigenvalues of $\widehat{V}_m$ are determined in the limit by the contamination among the first k slices by Theorem 4.1(b), we find that the space spanned by certain sub-collections of the

eigenvectors $\{\widehat{\eta}_{m,h}\}_{h=1}^k$ converges to the space spanned by certain sub-collections of the eigenvectors $\{\Sigma^{-1/2}\widetilde{\beta}_h\}_{h=1}^k$ and both collections of eigenvectors are determined by the number of observations $\{n_h\}_{h=1}^k$ in the first k slices. Namely, collections of slices have the same number of observations n_h correspond to collections of eigenvectors that converge.

Proof: Using the known μ, Σ values with the sample $(X, Y)_m^{n,k}$, the estimated covariance matrix $\widehat{V}_m$ can now be written in two parts:

$$\widehat{V}_m = \Sigma^{-1/2} \sum_{h=1}^k \widehat{p}_h(\overline{x}_{h,m} - \mu)(\overline{x}_{h,m} - \mu)^\top \Sigma^{-1/2} + \Sigma^{-1/2} \sum_{h=k+1}^H \widehat{p}_h(\overline{x}_h - \mu)(\overline{x}_h - \mu)^\top \Sigma^{-1/2}$$

with

$$\overline{x}_h = \frac{1}{n_h} \sum_{i:y_i \in \mathbb{I}_h} x_i, \qquad h = k+1, \ldots, H;$$

$$\overline{x}_{h,m} = \frac{1}{n_h} \widetilde{x}_{h,m} + \frac{1}{n_h} \sum_{i=2}^{n_h} x_i$$

$$\stackrel{(4.21)}{=} \frac{t_m}{n_h} \widetilde{\beta}_h + \frac{1}{n_h}\Big(\widetilde{v}_h + \sum_{i=2}^{n_h} x_i\Big) = \frac{t_m}{n_h} \widetilde{\beta}_h + v_h + \mu, \qquad h = 1, \ldots, k,$$

where $v_h = n_h^{-1}\big(\widetilde{v}_h + \sum_{i=2}^{n_h} x_i\big) - \mu$. With some algebra, we express

$$\widehat{V}_m = \sum_{h=1}^k \widehat{p}_h \Sigma^{-1/2} \Big(\frac{t_m}{n_h}\widetilde{\beta}_h + v_h\Big)\Big(\frac{t_m}{n_h}\widetilde{\beta}_h + v_h\Big)^\top \Sigma^{-1/2} + S$$

$$= \sum_{h=1}^k \Big(\frac{\widehat{p}_h}{n_h^2} t_m^2 \overline{\beta}_h \overline{\beta}_h^\top + \frac{\widehat{p}_h}{n_h} t_m \overline{\beta}_h v_h^\top \Sigma^{-1/2} + \frac{\widehat{p}_h}{n_h} t_m \Sigma^{-1/2} v_h \overline{\beta}_h^\top\Big) + S, \quad (4.23)$$

$$\overline{\beta}_h = \Sigma^{-1/2} \widetilde{\beta}_h, \qquad S = \Sigma^{-1/2}\Big(\sum_{h=1}^k \widehat{p}_h v_h v_h^\top + \sum_{h=k+1}^H \widehat{p}_h(\overline{x}_h - \mu)(\overline{x}_h - \mu)^\top\Big)\Sigma^{-1/2}.$$

Note that the term S does not depend on the contaminated observations. Because

$$\lim_{m\to\infty} \frac{1}{t_m^2}\Big(\sum_{h=1}^k \widehat{p}_h t_m \overline{\beta}_h v_h^\top \Sigma^{-1/2} + \widehat{p}_h t_m \Sigma^{-1/2} v_h \overline{\beta}_h^\top + S\Big) = 0_{p\times p},$$

and $\widehat{p}_h = n_h/n$, it holds that

$$\lim_{m\to\infty} \frac{\widehat{V}_m}{t_m^2} = \lim_{m\to\infty} \sum_{h=1}^k \frac{\widehat{p}_h}{n_h^2} \overline{\beta}_h \overline{\beta}_h^\top = \sum_{h=1}^k \frac{1}{n_h n}(\Sigma^{-1/2}\widetilde{\beta}_h)(\Sigma^{-1/2}\widetilde{\beta}_h)^\top.$$

Note that the vectors $\overline{\beta}_h = \Sigma^{-1/2}\widetilde{\beta}_h$, $h = 1, \ldots, k$, are orthonormal by (4.22). Therefore, parts(a)-(d) of Theorem 4.1 now follow directly from applying Lemma 4.1, noting that $\frac{1}{n_{(k)}} \leq \cdots \leq \frac{1}{n_{(1)}}$.

To prove part (e), write $p \times k$ matrices $\widehat{N}_m = \left[\widehat{\eta}_{m,1} \cdots \widehat{\eta}_{m,k}\right]Q_m$ and $\widetilde{B} = \left[\widetilde{\beta}_1 \cdots \widetilde{\beta}_k\right]$. We now aim to show that

$$\lim_{m \to \infty} (\widehat{N}_m^\top \Sigma^{-1} \widehat{N}_m)^{-1} = (\widetilde{B}^\top \Sigma^{-2} \widetilde{B})^{-1}. \tag{4.24}$$

From (4.24) and

$$\lim_{m \to \infty} \Sigma^{-1/2}\widehat{N}_m = \Sigma^{-1}\widetilde{B} \tag{4.25}$$

from part (d), the result in part (e) follows directly:

$$\begin{aligned}
\lim_{m \to \infty} P_{\Sigma^{-1/2}\widehat{N}_m} &= \lim_{m \to \infty} \Sigma^{-1/2}\widehat{N}_m(\widehat{N}_m^\top \Sigma^{-1}\widehat{N}_m)^{-1}\widehat{N}_m^\top \Sigma^{-1/2} \\
&= \Sigma^{-1}\widetilde{B}(\widetilde{B}^\top \Sigma^{-2}\widetilde{B})^{-1}\widetilde{B}^\top \Sigma^{-1} \\
&= P_{\Sigma^{-1}\widetilde{\mathcal{B}}}.
\end{aligned}$$

Because Σ^{-1} is symmetric and positive definite, we may write $\Sigma^{-1} = Q^\top D Q$, where Q is a $p \times p$ orthogonal matrix and D is a $p \times p$ positive diagonal matrix. Define a $p \times k$ matrix $\widehat{N}_m^* = Q\widehat{N}_m$ and note that the columns of $\widehat{N}_m^*$ are orthonormal vectors (i.e., $(\widehat{N}_m^*)^\top(\widehat{N}_m^*) = I_k$). To finish the proof, we use some properties of the Frobenius $\|A\|_{\mathbb{F}}$ and spectral $\|A\|_2$ norms of a symmetric $k \times k$ matrix A. These norms are given in Definition 2.8 and (2.8), respectively. From Lemma 7.4, we have

$$\|A\|_2 \leq \|A\|_{\mathbb{F}} \leq \sqrt{k} \cdot \|A\|_2.$$

In addition, if B is a $k \times k$ matrix, then $\|AB\|_{\mathbb{F}} \leq k \cdot \|A\|_{\mathbb{F}}\|B\|_{\mathbb{F}}$. Using these properties, we let $C_m = (\widehat{N}_m^\top \Sigma^{-1}\widehat{N}_m)^{-1} = [(\widehat{N}_m^*)^\top D (\widehat{N}_m^*)]^{-1}$ and $C = (\widetilde{B}^\top \Sigma^{-2}\widetilde{B})^{-1}$ and write

$$\begin{aligned}
\mathrm{tr}\left[(C_m - C)^\top(C_m - C)\right] &= \|C_m - C\|_{\mathbb{F}}^2 \\
&= \|C_m(C_m^{-1} - C^{-1})C\|_{\mathbb{F}}^2 \\
&\leq k \cdot \|C_m\|_2^2 \cdot \|C_m^{-1} - C^{-1}\|_{\mathbb{F}}^2 \cdot \|C\|_{\mathbb{F}}^2 \\
&= k \cdot \|C\|_{\mathbb{F}}^2 \cdot \|D^{-1}\|_2^2 \cdot \mathrm{tr}\left[(C_m^{-1} - C^{-1})^\top(C_m^{-1} - C^{-1})\right],
\end{aligned}$$

where above, because C_m is symmetric,

$$
\begin{aligned}
\|C_m\|_2^2 &= \text{largest eigenvalue of } C_m C_m^\top \\
&= \text{largest eigenvalue of } \left[(C_m^{-1})(C_m^{-1}) \right]^{-1} \\
&= \text{largest eigenvalue of } \left[(\widehat{N}_m^*)^\top D(\widehat{N}_m^*)(\widehat{N}_m^*)^\top D(\widehat{N}_m^*) \right]^{-1} \\
&= \text{largest eigenvalue of } \left[(\widehat{N}_m^*)^\top D^2(\widehat{N}_m^*) \right]^{-1} \\
&= \text{largest eigenvalue of } [D^2]^{-1} \\
&= \|D^{-1}\|_2^2.
\end{aligned}
$$

By (4.25), $\lim_{m\to\infty}(C_m^{-1} - C^{-1}) = 0_{k\times k}$ so that it follows that $\lim_{m\to\infty} \text{tr}\big[(C_m - C)^\top (C_m - C)\big] = 0$ or, equivalently, that $\lim_{m\to\infty}(C_m - C)^\top(C_m - C) = 0_{k\times k}$. This proves (4.24) and part (e). $\square$

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

Corollary 4.1 *For a given $1 \leq \mathcal{K} < p$, suppose the SIR procedure seeks to estimate a $\mathcal{K}$-dimensional subspace of $\mathbb{R}^p$ based on $H \geq \min\{\mathcal{K}, p - \mathcal{K}\}$ data slices with a size n data sample $(X, Y)^n$ and known values of $\mathrm{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$. Then,*

(i) there exists a sequence $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, of contaminated data sets and associated subspace estimates $\widehat{\mathcal{B}}_{m,k}$, found by replacing $k = \min\{\mathcal{K}, p-\mathcal{K}\}$ observations in $(X, Y)^n$, where

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

where $\widehat{\mathcal{B}}$ is the subspace estimate based on $(X, Y)^n$.
(ii) the finite sample breakdown point of SIR, under Definition 4.1, satisfies

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

Remark. We make a few comments on the nature of the proof and the contamination scheme used. Uncontaminated data $(X, Y)^n$ produces a $\mathcal{K}-$dimensional estimate $\widehat{\mathcal{B}} \subset$

$\mathbb{R}^p$ of the reduction space, when the dimension $\mathcal{K}$ of reduction is known. If $\widehat{\mathcal{B}}_{m,k} \subset \mathbb{R}^p$ represents a corrupted $\mathcal{K}-$dimensional estimate of the reduction subspace, based on replacing k observations in $(X, Y)^n$ under the contamination scheme in Section 4.3.1, then the maximal distance $\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k})$ between the corrupted and uncorrupted subspace estimates is given in (4.1). There are the two possible cases for this distance, depending on whether $2\mathcal{K} \leq p$ or $2\mathcal{K} > p$.

If $2\mathcal{K} \leq p$, the subspace estimates $\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k}$ are maximally distant under $\mathbb{F}$ if these estimates are orthogonal. In this case, by appropriately replacing $k = \mathcal{K}$ observations in the first $\mathcal{K}$ slices, we can obtain a contaminated estimate $\widehat{\mathcal{B}}_{m,k=\mathcal{K}}$ that is arbitrarily close to being orthogonal to the uncontaminated estimate $\widehat{\mathcal{B}}$ (in the sense that $P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} \approx 0_{p \times p}$).

If $2\mathcal{K} > p$, the subspace estimates $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_{m,k}$ must have a non-empty intersection that is at least of dimension $2\mathcal{K} - p$ by Lemma 2.2. In this case, subspaces $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_{m,k}$ are maximally distant under $\mathbb{F}$ if the intersection $\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k}$ is as minimal as possible (e.g., $\dim(\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k}) = 2\mathcal{K} - p$) and, after removing the intersection, the remaining parts of both subspaces $\widehat{\mathcal{B}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp$ and $\widehat{\mathcal{B}}_{m,k} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp$ are orthogonal. Here we can replace $k = p - \mathcal{K}$ observations to obtain a contaminated sample and an estimate $\widehat{\mathcal{B}}_{m,k=p-\mathcal{K}}$ for which $\dim(\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}) \approx 2\mathcal{K} - p$ and $P_{\widehat{\mathcal{B}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp} \approx 0_{p \times p}$, ensuring enough contamination to force a breakdown of the dimension reduction procedure.

Proof. Let $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ denote the $\mathcal{K}$-dimensional reduction subspace determined from the uncontaminated data; $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ is spanned by the scaled eigenvectors $\widehat{\beta}_1 = \Sigma^{-1/2}\widehat{\eta}_1, \ldots, \widehat{\beta}_{\mathcal{K}} = \Sigma^{-1/2}\widehat{\eta}_{\mathcal{K}}$ associated with the $\mathcal{K}$ largest eigenvalues of (4.20) computed with $(X, Y)^n$. To establish a bound on $\epsilon_{fsbp,\mathcal{K}}$, we consider two possible cases: $2\mathcal{K} \leq p$ or $2\mathcal{K} > p$. For each case, we examine the finite sample breakdown point under different contamination schemes which satisfy (4.21) and (4.22).

CASE I: $2\mathcal{K} \leq p$. Here we use the contamination routine described in Section 4.3.1 to obtain a contaminated sample upon replacing $\mathcal{K}$ observations in $(X, Y)^n$. Choose

orthonormal $\overline{\beta}_1, \ldots, \overline{\beta}_{\mathcal{K}} \in \Sigma^{1/2}\widehat{\mathcal{B}}^{\perp} = \{\Sigma^{1/2}v : v \in \widehat{\mathcal{B}}^{\perp}\} \subset \mathbb{R}^p$, which is possible because $\dim(\Sigma^{1/2}\widehat{\mathcal{B}}^{\perp}) = \dim(\widehat{\mathcal{B}}^{\perp}) = p - \dim(\widehat{\mathcal{B}}) = p - \mathcal{K} \geq \mathcal{K}$. Following Section 4.3.1, replace the first observation in the first $\mathcal{K}$ slices with contaminated values $(\widetilde{x}_{1_h,m}, \widetilde{y}_{1_h}) = (\widetilde{x}_{1_h}, y_{1_h})$ to obtain a sequence of contaminated samples $(X,Y)_m^{n,\mathcal{K}}$ for $m \in \mathbb{N}$, where $\widetilde{x}_{1_h,m} = t_m \widetilde{\beta}_h$ with $\widetilde{\beta}_h = \Sigma^{1/2}\overline{\beta}_h$, for $h = 1, \ldots, \mathcal{K}$. Note that the vectors $\{\widetilde{\beta}_h\}_{h=1}^{\mathcal{K}}$ satisfy (4.22).

After computing a covariance estimate $\widehat{V}_m$ based on $(X,Y)_m^{n,\mathcal{K}}$ in (4.20), an application of the SIR method to $(X,Y)_m^{n,\mathcal{K}}$ results in an estimated $\mathcal{K}$-dimensional reduction subspace $\widehat{\mathcal{B}}_{m,\mathcal{K}} \subset \mathbb{R}^p$ spanned by the scaled eigenvectors $\widehat{\beta}_{m,1} = \Sigma^{-1/2}\widehat{\eta}_{m,1}, \ldots, \widehat{\beta}_{m,\mathcal{K}} = \Sigma^{-1/2}\widehat{\eta}_{m,\mathcal{K}}$ of $\widehat{V}_m$ associated with the $\mathcal{K}$ largest eigenvalues of $\widehat{V}_m$. Let $\Sigma^{-1}\widetilde{\mathcal{B}} \subset \mathbb{R}^p$ denote the subspace spanned by $\Sigma^{-1}\widetilde{\beta}_1 = \Sigma^{-1/2}\overline{\beta}_1, \ldots, \Sigma^{-1}\widetilde{\beta}_{\mathcal{K}} = \Sigma^{-1/2}\overline{\beta}_{\mathcal{K}}$. It follows from Theorem 4.1 (e) (with $k = \mathcal{K}$) that

$$\lim_{m \to \infty} P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} = P_{\Sigma^{-1}\widetilde{\mathcal{B}}}$$

and so

$$
\begin{aligned}
\lim_{m \to \infty} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,\mathcal{K}}) &= \lim_{m \to \infty} \sqrt{2\mathcal{K} - 2\mathrm{tr}(P_{\widehat{\mathcal{B}}}P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}})} \\
&= \sqrt{2\mathcal{K} - 2\mathrm{tr}(P_{\widehat{\mathcal{B}}}P_{\Sigma^{-1}\widetilde{\mathcal{B}}})} = \mathbb{F}(\widehat{\mathcal{B}}, \Sigma^{-1}\widetilde{\mathcal{B}})
\end{aligned}
$$

by the continuity of matrix trace and multiplication operations in (2.2). Because $\Sigma^{-1}\widetilde{\mathcal{B}} \subset \widehat{\mathcal{B}}^{\perp}$, it holds that $P_{\Sigma^{-1}\widetilde{\mathcal{B}}}P_{\widehat{\mathcal{B}}} = 0_{p \times p}$ by orthogonality. From this and $2\mathcal{K} \leq p$, the subspaces $\widehat{\mathcal{B}}$ and $\Sigma^{-1}\widetilde{\mathcal{B}}$ are maximally distant $\mathcal{K}$-dimensional subspaces under $\mathbb{F}$ by (4.1), namely $\mathbb{F}(\widehat{\mathcal{B}}, \Sigma^{-1}\widetilde{\mathcal{B}}) = \sqrt{2\mathcal{K}}$. By Definition 4.1, we conclude

$$\epsilon_{fsbp,\mathcal{K}}((X,Y)^n, \mathbb{F}, \mathcal{K}) \leq \frac{\mathcal{K}}{n} = \frac{\min\{\mathcal{K}, p - \mathcal{K}\}}{n}.$$

CASE II: $2\mathcal{K} > p$. Again we use the contamination scheme from Section 4.3.1, but replace $k = p - \mathcal{K}$ observations in $(X,Y)^n$ for a contaminated sample. Choose orthonormal $\overline{\beta}_1, \ldots, \overline{\beta}_{p-\mathcal{K}}$ vectors which span $\Sigma^{1/2}\widehat{\mathcal{B}}^{\perp} \subset \mathbb{R}^p$, noting $\dim(\Sigma^{1/2}\widehat{\mathcal{B}}^{\perp}) = \dim(\widehat{\mathcal{B}}^{\perp}) = p - \mathcal{K}$. To obtain a sequence of contaminated samples $(X,Y)_m^{n,p-\mathcal{K}}$, $m \in \mathbb{N}$, we replace the first observation in the first $k = p - \mathcal{K}$ slices with contaminated values $(\widetilde{x}_{1_h,m}, \widetilde{y}_{1_h}) = (\widetilde{x}_{1_h}, y_{1_h})$ where $\widetilde{x}_{1_h,m} = t_m \widetilde{\beta}_h$ with $\widetilde{\beta}_h = \Sigma^{1/2}\overline{\beta}_h$, for $h = 1, \ldots, p - \mathcal{K}$.

The Sir procedure, applied to $(X, Y)_m^{n, p-\mathcal{K}}$, again results in an estimated $\mathcal{K}$-dimensional reduction subspace $\widehat{\mathcal{B}}_{m, p-\mathcal{K}} \subset \mathbb{R}^p$ spanned by the $\Sigma^{-1/2}$-scaled eigenvectors $\widehat{\beta}_{m,1}, \ldots, \widehat{\beta}_{m,\mathcal{K}}$ of a covariance estimate $\widehat{V}_m$ (associated with the $\mathcal{K}$ largest eigenvalues $\widehat{\lambda}_{m,1} \geq \cdots \geq \widehat{\lambda}_{m,\mathcal{K}}$ of $\widehat{V}_m$). Write $\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}$ and $\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(2)}$ to denote the $\mathbb{R}^p$-subspaces spanned by vectors $\{\widehat{\beta}_{m,h}\}_{h=1}^{p-\mathcal{K}}$ and $\{\widehat{\beta}_{m,h}\}_{h=p-\mathcal{K}+1}^{\mathcal{K}}$, respectively. By orthogonality, we may write

$$P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} = P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}} + P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(2)}}.$$

Because $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_{m,p-\mathcal{K}}$ are $\mathcal{K}-$dimensional subspaces of $\mathbb{R}^p$, the dimension of the intersection $\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}$ is at least $2\mathcal{K}-p$ by Lemma 2.2, for all $m \in \mathbb{N}$. Because the projection matrix for $\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}})^\perp$ is $P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}})^\perp} = P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} - P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}$ and $P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}} = P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}$ from $\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}} \subset \widehat{\mathcal{B}}$, it follows that for all $m \in \mathbb{N}$:

$$
\begin{aligned}
\operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) &= \operatorname{tr}\left[P_{\widehat{\mathcal{B}}}\big((P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} - P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) + P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}\big)\right] \\
&= \operatorname{tr}\left[P_{\widehat{\mathcal{B}}}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}})^\perp} + P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}})\right] \\
&= \operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}})^\perp}) + \operatorname{tr}(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) \\
&\geq \operatorname{tr}(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) \geq 2\mathcal{K} - p \qquad (4.26)
\end{aligned}
$$

using Lemma 7.3. We note as well that

$$
\begin{aligned}
\operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(2)}}) &\leq \operatorname{rank}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(2)}}) \\
&= \operatorname{tr}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(2)}}) \\
&= \operatorname{tr}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} - P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}}) \\
&= \operatorname{tr}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) - \operatorname{tr}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}}) \qquad (4.27) \\
&= \operatorname{rank}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) - \operatorname{rank}(P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}}) = 2\mathcal{K} - p.
\end{aligned}
$$

Let $\Sigma^{-1}\widetilde{\mathcal{B}} \subset \mathbb{R}^p$ denote the subspace spanned by $\Sigma^{-1}\widetilde{\beta}_1 = \Sigma^{-1/2}\overline{\beta}_1, \ldots, \Sigma^{-1}\widetilde{\beta}_{p-\mathcal{K}} = \Sigma^{-1/2}\overline{\beta}_{p-\mathcal{K}}$. From Theorem 4.1 (e) (with $k = p - \mathcal{K}$), it follows directly that

$$\lim_{m \to \infty} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}^{(1)}} = P_{\Sigma^{-1}\widetilde{\mathcal{B}}}.$$

By this and (4.26)-(4.28), we find

$$
\begin{aligned}
2\mathcal{K} - p \ &\leq\ \lim_{m \to \infty} \operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) \\
&=\ \lim_{m \to \infty} \operatorname{tr}\big[P_{\widehat{\mathcal{B}}} (P_{\widehat{\mathcal{B}}^{(1)}_{m,p-\mathcal{K}}} + P_{\widehat{\mathcal{B}}^{(2)}_{m,p-\mathcal{K}}}) \big] \\
&\leq\ \lim_{m \to \infty} \operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}^{(1)}_{m,p-\mathcal{K}}}) + 2\mathcal{K} - p \\
&=\ \lim_{m \to \infty} \operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\Sigma^{-1}\widetilde{\mathcal{B}}}) + 2\mathcal{K} - p \\
&=\ 2\mathcal{K} - p,
\end{aligned}
$$

using above $P_{\widehat{\mathcal{B}}} P_{\Sigma^{-1}\widetilde{\mathcal{B}}} = 0_{p \times p}$ by orthogonality since $\Sigma^{-1}\widetilde{\mathcal{B}} = \widehat{\mathcal{B}}^{\perp}$. Hence, $\lim_{m \to \infty} \operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) = 2\mathcal{K} - p$ and so we finally derive

$$
\lim_{m \to \infty} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,p-\mathcal{K}}) = \lim_{m \to \infty} \sqrt{2\mathcal{K} - 2\operatorname{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}})} = \sqrt{2(p - \mathcal{K})}.
$$

The limiting value $\sqrt{2(p - \mathcal{K})}$ of $\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,p-\mathcal{K}})$ is the maximal distance between two $\mathcal{K}$-dimensional subspaces of $\mathbb{R}^p$ under $\mathbb{F}$ by (4.1) in the considered case $2\mathcal{K} > p$. By Definition 4.1, we may conclude

$$
\epsilon_{fsbp,\mathcal{K}}((X,Y)^n, \mathbb{F}, \mathcal{K}) \ \leq\ \frac{p - \mathcal{K}}{n} = \frac{\min\{\mathcal{K}, p - \mathcal{K}\}}{n}. \qquad \square
$$

4.3.2 $\mathcal{K}$ known $- \Sigma$ unknown case

Here we establish a bound on the finite sample breakdown point of SIR in the case that the dimension $1 \leq \mathcal{K} < p$ of the reduction subspace is known, but the components of the covariance structure $\operatorname{Cov}(X) = \Sigma$, $\mathrm{E}(X) = \mu$ are unknown and require estimation.

As in Section 4.3.1, let $(X,Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ denote an uncontaminated sample of size n. We estimate the unknown mean μ and variance Σ with $\overline{x} = \sum_{i=1}^n x_i/n$ and $\widehat{\Sigma} = \sum_{i=1}^n (x_i - \overline{x})(x_i - \overline{x})^\top/n$, respectively. Based on $(X,Y)^n$, a SIR-type procedure results in a $\mathcal{K}-$dimensional reduction subspace estimate $\widehat{\mathcal{B}}$ based on $\widehat{\beta}_1 = \widehat{\Sigma}^{-1/2}\widehat{\eta}_1, \ldots, \widehat{\beta}_\mathcal{K} = \widehat{\Sigma}^{-1/2}\widehat{\eta}_\mathcal{K}$, where $\widehat{\eta}_1, \ldots, \widehat{\eta}_\mathcal{K}$ are the eigenvectors corresponding to the $\mathcal{K}$ largest eigenvectors $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_\mathcal{K}$ of

$$
\widehat{V} = \widehat{\Sigma}^{-1/2} \sum_{h=1}^H \widehat{p}_h (\overline{x}_h - \overline{x})(\overline{x}_h - \overline{x})^\top \widehat{\Sigma}^{-1/2}. \tag{4.28}
$$

Equivalently, $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$ can be determined as the eigenvectors corresponding to the $\mathcal{K}$ largest eigenvectors of

$$\widetilde{V} = \widehat{\Sigma}^{-1} \sum_{h=1}^{H} \widehat{p}_h (\overline{x}_h - \overline{x})(\overline{x}_h - \overline{x})^{\top}, \tag{4.29}$$

since $\widehat{V}\widehat{\eta}_i = \widehat{\lambda}_i \widehat{\eta}_i$ if and only if $\widetilde{V}\widehat{\beta}_i = \widehat{\lambda}_i \widehat{\beta}_i$, $i = 1, \ldots, \mathcal{K}$.

In our subsequent study of the breakdown properties of SIR, we use a contamination scheme similar to Section 4.3.1 to obtain a contaminated sample $(X, Y)_m^{n,k}$ based on replacing k observations among the first k slices $\mathbb{I}_i$, $i = 1, \ldots, H$.

To establish a bound on the breakdown point of SIR, we will make use of the following results.

Lemma 4.2 *Let M_0 be a $p \times p$ positive definite matrix and let $\{t_m\}_{m=1}^{\infty}$ be a positive, increasing real sequence such that $t_m \to \infty$ as $m \to \infty$. For $i = 1, \ldots, k < p$, let $\beta_i \neq 0_p, u_i \in \mathbb{R}^p$, $c_i > 0 \in \mathbb{R}$ such that $\{\beta_i\}_{i=1}^{k}$ are linearly independent; define $\beta_{m,i} = \beta_i + (u_i/t_m)$ and let*

$$M_{m,i} = M_0 + t_m^2 \sum_{j=1}^{i} c_j \beta_{m,j} \beta_{m,j}^{\top}, \qquad m \in \mathbb{N}.$$

(a) Then, for each $i = 1, \ldots, k$ and $m \in \mathbb{N}$, the matrix $M_{m,i}$ is positive definite with an inverse given by

$$M_{m,i}^{-1} = M_{m,i-1}^{-1} - \frac{c_i \cdot t_m^2 \cdot M_{m,i-1}^{-1} \beta_{m,i} \beta_{m,i}^{\top} M_{m,i-1}^{-1}}{1 + c_i \cdot t_m^2 \cdot \beta_{m,i}^{\top} M_{m,i-1}^{-1} \beta_{m,i}} \tag{4.30}$$

where $M_{m,0} = M_0$ for $m \geq 1$. Furthermore, for $i = 1, \ldots, k$:

$$\begin{aligned} M_{\infty,i}^{-1} &\equiv \lim_{m \to \infty} M_{m,i}^{-1} = M_0^{-1/2}\left(\mathrm{I}_p - P_{M_0^{-1/2}B_i}\right)M_0^{-1/2}, \\ \lim_{m \to \infty} t_m^2 \cdot M_{m,i}^{-1}\beta_{m,i} &= \frac{M_{\infty,i-1}^{-1}\beta_i}{c_i \cdot \beta_i^{\top} M_{\infty,i-1}^{-1}\beta_i}, \end{aligned} \tag{4.31}$$

where $B_i = [\beta_1 \cdots \beta_i]$ represents a $p \times i$ matrix and $M_{\infty,0}^{-1} \equiv M_0^{-1}$.

(b) Let $\overline{\beta}_m = \sum_{i=1}^{k} \beta_{m,i}/k$ and $M_m = M_{m,k} + t_m^2 \sum_{i=1}^{k} \widetilde{c}_i(\beta_{m,i} - \overline{\beta}_m)(\beta_{m,i} - \overline{\beta}_m)^{\top}$ for $\widetilde{c}_i > 0$, $i = 1, \ldots, k$. Then, M_m is positive definite and satisfies

$$\lim_{m \to \infty} M_m^{-1} = M_0^{-1/2}\left(\mathrm{I}_p - P_{M_0^{-1/2}B_k}\right)M_0^{-1/2}$$

with $\sup\limits_{m} t_m^2 \|M_m^{-1}\beta_{m,i}\| < \infty$ *for* $i = 1, \ldots, k$.

Remark. The above results imply that, because of the contamination/pertubation of M_0 to M_m (or $M_{m,k}$), the limit of M_m^{-1} is *not* positive definite. In fact, the contamination vectors $\{\beta_i\}_{i=1}^k$ span the nullspace of the limiting inverse matrix $\lim_{m\to\infty} M_m^{-1}$. That is, for $v \in \mathbb{R}^p$, we find $(\lim_{m\to\infty} M_{m,k}^{-1})v = 0_p$ if and only if $v \in \mathrm{span}(\{\beta_i\}_{i=1}^k)$.

Proof. We prove part (a) of the lemma first. Given the form of $M_{\infty,i-1}^{-1}$ in (4.31) (with $M_{\infty,0}^{-1} \equiv M_0^{-1}$ and $P_{M_0^{-1/2}B_0} \equiv 0_{p\times p}$), we note that

$$
\begin{aligned}
\beta_i^\top M_{\infty,i-1}^{-1}\beta_i > 0 &\iff \left(\mathrm{I}_p - P_{M_0^{-1/2}B_{i-1}}\right)M_0^{-1/2}\beta_i \neq 0_p \\
&\iff M_0^{-1/2}\beta_i \notin \mathrm{span}(\{M_0^{-1/2}\beta_j\}_{j=1}^{i-1}) \qquad (4.32) \\
&\iff \beta_i \notin \mathrm{span}(\{\beta_j\}_{j=1}^{i-1}), \quad i = 1, \ldots, k.
\end{aligned}
$$

Because $\{\beta_i\}_{i=1}^k$ are linearly independent, we have that $\beta_i \notin \mathrm{span}(\{\beta_j\}_{j=1}^{i-1})$ for each $i = 1, \ldots, k$ and the denominator of (4.31) is well-defined, if the form of each $M_{\infty,i}^{-1}$ holds as claimed.

We use a proof by induction. Consider $i = 1$ first. Then, $M_{m,1}$ is positive definite from the fact that for $v \neq 0_p \in \mathbb{R}^p$, $i = 1$,

$$
v^\top M_{m,i} v = v^\top M_{m,i-1} v + t_m^2 \cdot c_i \cdot (v^\top \beta_{m,i})^2 \geq v^\top M_{m,i-1} v > 0 \qquad (4.33)
$$

by the positive definiteness of $M_{m,i-1}$. One can verify directly that $M_{m,1}^{-1}$ in (4.30) satisfies $M_{m,1} M_{m,1}^{-1} = \mathrm{I}_p$. The limiting properties of $M_{m,1}^{-1}$ follow immediately upon noting that, for $i = 1$,

$$
M_{m,i}^{-1}\beta_{m,i} = \frac{M_{m,i-1}^{-1}\beta_{m,i}}{1 + c_i \cdot t_m^2 \cdot \beta_{m,i}^\top M_{m,i-1}^{-1}\beta_{m,i}} \qquad (4.34)
$$

and $P_{M_0^{-1/2}B_1} = M_0^{-1/2}\beta_1\beta_1^\top M_0^{-1/2}/(\beta_1^\top M_0^{-1}\beta_1)$.

We show now that the claim holds for a given $i = 2, \ldots, k$, under the induction assumption that $M_{m,i-1}$ is positive definite with inverse $M_{m,i-1}^{-1}$ that has the two claimed limiting properties. If $M_{m,i-1}$ is positive definite, $M_{m,i}$ is positive definite by the same

argument as in (4.33) and, since multiplying the matrix (4.30) by $M_{m,i}$ yields the identity matrix I_p, the matrix in (4.30) must be $M_{m,i}^{-1}$. Because $M_{\infty,i-1}^{-1} \equiv \lim_{m\to\infty} M_{m,i-1}$ exists by the induction hypothesis with form given by (4.31), $\beta_i M_{\infty,i-1}^{-1} \beta_i^\top > 0$ and $\left(I_p - P_{M_0^{-1/2}B_{i-1}}\right) M_0^{-1/2}\beta_i \neq 0_p$ hold by (4.32) under the induction assumption, and $\lim_{m\to\infty} \beta_{m,i} = \beta_i$, we can deduce that

$$
\begin{aligned}
\lim_{m\to\infty} M_{m,i}^{-1} &= M_{\infty,i-1}^{-1} - \frac{M_{\infty,i-1}^{-1}\beta_i\beta_i^\top M_{\infty,i-1}^{-1}}{\beta_i^\top M_{\infty,i-1}^{-1}\beta_i} \\
&= M_{\infty,i-1}^{-1} - M_0^{-1/2} P_{(I_p - P_{M_0^{-1/2}B_{i-1}})M_0^{-1/2}\beta_i} M_0^{-1/2} \\
&= M_0^{-1/2}\left(I_p - P_{M_0^{-1/2}B_{i-1}} - P_{(I_p - P_{M_0^{-1/2}B_{i-1}})M_0^{-1/2}\beta_i}\right) M_0^{-1/2} \\
&= M_0^{-1/2}\left(I_p - P_{M_0^{-1/2}B_i}\right) M_0^{-1/2}.
\end{aligned}
$$

The last equality above follows from the fact that $P_{M_0^{-1/2}B_{i-1}} + P_{(I_p - P_{M_0^{-1/2}B_{i-1}})M_0^{-1/2}\beta_i}$ is the orthogonal projection matrix for the $\mathbb{R}^p$-subspace spanned by vectors $\{M_0^{-1/2}B_{i-1}, (I_p - P_{M_0^{-1/2}B_{i-1}})M_0^{-1/2}\beta_i\}$ (i.e., the sum of the two projection matrices is symmetric and idempotent since $P_{M_0^{-1/2}B_{i-1}} P_{(I_p - P_{M_0^{-1/2}B_{i-1}})M_0^{-1/2}\beta_i} = 0_{p\times p}$ holds by orthogonality). Because the two collections of vectors $\{M_0^{-1/2}B_{i-1}, (I_p - P_{M_0^{-1/2}B_{i-1}})\beta_i\}$ and $M_0^{-1/2}B_i$ span the same subspace of $\mathbb{R}^p$, it must be the case that $P_{M_0^{-1/2}B_i} = P_{M_0^{-1/2}B_{i-1}} + P_{(\mathbb{I}_p - P_{M_0^{-1/2}B_{i-1}})\beta_i}$ by the uniqueness of projection matrices. Likewise, we find (4.34) holds for i using the formula of $M_{m,i}^{-1}$ so that $\lim_{m\to\infty} t_m^2 \cdot M_{\infty,i}^{-1}\beta_{m,i}$ again follows by the induction assumption. This completes the proof by induction of part (a).

To establish part (b), define $\widetilde{M}_{m,0} = M_{m,k}$ and $\widetilde{\beta}_{m,i} = \beta_{m,i} - \overline{\beta}_m$,

$$
\widetilde{M}_{m,i} = \widetilde{M}_{m,0} + t_m^2 \sum_{j=1}^{i} \widetilde{c}_j \widetilde{\beta}_{m,j}\widetilde{\beta}_{m,j}^\top, \quad i = 1, \ldots, k.
$$

Note that each $\widetilde{M}_{m,i}$, $i = 1, \ldots, k$, is positive definite (e.g., $v^\top \widetilde{M}_{m,i} v > 0$, $v \neq 0_p \in \mathbb{R}^p$) because $M_{m,k} = \widetilde{M}_{m,0}$ is positive definite from part (a) and $\widetilde{M}_{m,i} - \widetilde{M}_{m,0}$ is clearly nonnegative definite. For each $i = 1, \ldots, k$, the inverse $\widetilde{M}_{m,i}^{-1}$ satisfies

$$
\widetilde{M}_{m,i}^{-1} = \widetilde{M}_{m,i-1}^{-1} - \frac{\widetilde{c}_i \cdot t_m^2 \cdot \widetilde{M}_{m,i-1}^{-1}\widetilde{\beta}_{m,i}\widetilde{\beta}_{m,i}^\top \widetilde{M}_{m,i-1}^{-1}}{1 + \widetilde{c}_i \cdot t_m^2 \cdot \widetilde{\beta}_{m,i}^\top \widetilde{M}_{m,i-1}^{-1}\widetilde{\beta}_{m,i}} \tag{4.35}
$$

upon replacing $\big\{ M_{m,i}, M_{m,i-1}, \beta_{m,i}, c_i \big\}$ with counterparts $\big\{ \widetilde{M}_{m,i}, \widetilde{M}_{m,i-1}, \widetilde{\beta}_{m,i}, \widetilde{c}_i \big\}$ in (4.30). Because $M_m = \widetilde{M}_{m,k}$, it suffices to show that for each $i = 0, 1, \ldots, k$,

$$
\lim_{m \to \infty} \widetilde{M}_{m,i}^{-1} = M_0^{-1/2} \Big(\mathrm{I}_p - P_{M_0^{-1/2} B_k} \Big) M_0^{-1/2} \equiv M_{\infty,k}^{-1};
$$

$$
\sup_m t_m^2 \cdot \| \widetilde{M}_{m,i}^{-1} \beta_{m,j} \| < \infty, \quad j = 1, \ldots, k. \tag{4.36}
$$

Again we proceed with an induction argument to show (4.36). For $i = 1, \ldots, k$, let B_{i-} denote the $p \times (k-1)$ matrix that results from removing the column β_i from B_k. For any $j = 1, \ldots, k$, we can rewrite $M_{m,k} = (M_0 + t_m^2 \sum_{\substack{\ell=1 \\ \ell \neq j}}^{k} c_\ell \beta_{m,\ell} \beta_{m,\ell}^\top) + c_j \beta_{m,j} \beta_{m,j}^\top$ so that the jth term, involving $c_j \beta_{m,j} \beta_{m,j}^\top$, appears as the last (i.e., kth) term in the sum $M_{m,k}$. Treating the jth term $c_j \beta_{m,j} \beta_{m,j}^\top$ as the (final) kth term in the sum $M_{m,k}$, it follows from the form of inverses in part (a) and (4.31) that for any $j = 1, \ldots, k$,

$$
\lim_{m \to \infty} t_m^2 \cdot M_{m,k}^{-1} \beta_{m,j} = \frac{M_0^{-1/2} \Big(\mathrm{I}_p - P_{M_0^{-1/2} B_{j-}} \Big) M_0^{-1/2} \beta_j}{c_j \cdot \beta_j^\top M_0^{-1/2} \Big(\mathrm{I}_p - P_{M_0^{-1/2} B_{j-}} \Big) M_0^{-1/2} \beta_j}
$$

$$
\neq 0_p, \quad (j = 1, \ldots, k),
$$

where the last inequality follows from (4.32). Hence, (4.36) holds for $i = 0$ with $\widetilde{M}_{m,0} = M_{m,k}$.

Now assuming (4.36) holds for some $i \in \{0, 1, \ldots, k-1\}$, we show (4.36) holds for $i+1$. Under the induction assumption and

$$
\widetilde{\beta}_{m,i+1} = - \sum_{\substack{j=1 \\ j \neq i+1}}^{k} \frac{1}{k} \beta_{m,j} + \Big(1 - \frac{1}{k} \Big) \beta_{m,i+1},
$$

we find

$$
\sup_m t_m^2 \cdot \| \widetilde{M}_{m,i}^{-1} \widetilde{\beta}_{m,i+1} \| \leq \sum_{j=1}^{k} \max \Big\{ 1 - \frac{1}{k}, \frac{1}{k} \Big\} \cdot \sup_m t_m^2 \cdot \| \widetilde{M}_{m,i}^{-1} \beta_{m,j} \| < \infty, \tag{4.37}
$$

which implies

$$
\lim_{m \to \infty} \widetilde{M}_{m,i}^{-1} \widetilde{\beta}_{m,i+1} = 0_p, \quad 1 \leq \sup_m \{ 1 + \widetilde{c}_{i+1} \cdot t_m^2 \cdot \widetilde{\beta}_{m,i+1}^\top \widetilde{M}_{m,i}^{-1} \widetilde{\beta}_{m,i+1} \} \leq C < \infty \tag{4.38}
$$

for some $C > 0$ so that

$$
\lim_{m \to \infty} \frac{\widetilde{c}_{i+1} \cdot t_m^2 \cdot \widetilde{M}_{m,i}^{-1} \widetilde{\beta}_{m,i+1} \widetilde{\beta}_{m,i+1}^\top \widetilde{M}_{m,i}^{-1}}{1 + \widetilde{c}_{i+1} \cdot t_m^2 \cdot \widetilde{\beta}_{m,i+1}^\top \widetilde{M}_{m,i}^{-1} \widetilde{\beta}_{m,i+1}} = 0_{p \times p}.
$$

From this last limit and $\lim_{m\to\infty}\widetilde{M}_{m,i}^{-1}=M_{\infty,k}^{-1}$ by the induction assumption, we find $\lim_{m\to\infty}\widetilde{M}_{m,i+1}^{-1}=M_{\infty,k}^{-1}$ from (4.35). Finally, using the inverse in (4.35), recognizing that (4.36) holds for i under the induction assumption, and applying (4.37)-(4.38), we find that for any $j=1,\ldots,k$:

$$
\begin{aligned}
\sup_m t_m^2\cdot\|\widetilde{M}_{m,i+1}^{-1}\beta_{m,j}\| \;\leq\;& \sup_m t_m^2\cdot\|\widetilde{M}_{m,i}^{-1}\beta_{m,j}\| \\
&+\frac{\widetilde{c}_{i+1}}{C}\cdot\left(\sup_m t_m^2\cdot\|\widetilde{M}_{m,i}^{-1}\widetilde{\beta}_{m,i+1}\|\right)\cdot\left(\sup_m t_m^2\cdot\widetilde{\beta}_{m,i+1}^{\top}\widetilde{M}_{m,i}^{-1}\beta_{m,j}\right) \\
\leq\;& \sup_m t_m^2\cdot\|\widetilde{M}_{m,i}^{-1}\beta_{m,j}\| \\
&+\frac{\widetilde{c}_{i+1}}{C}\cdot\left(\sup_m t_m^2\cdot\|\widetilde{M}_{m,i}^{-1}\widetilde{\beta}_{m,i+1}\|\right)^2\cdot\sup_m\|\beta_{m,j}\| \\
<\;& \infty.
\end{aligned}
$$

Hence, we find (4.36) holds for $i+1$ which completes the induction proof of part (b). $\square$

The following is a generalization of Lemma 4.18 of Hilker (1997).

Lemma 4.3 *Let $\{M_m\}_{m=1}^{\infty}$ be a sequence of $p\times p$ matrices, each having p real eigenvalues. Suppose $\lim_{m\to\infty}M_m=W\sum_{i=1}^{k}u_i v_i^{\top}$, where W is a $p\times p$ matrix and $u_i,v_i\neq 0_p\in\mathbb{R}^p$, $W^{\top}v_i=0_p$, $i=1,\ldots,k$.*

Let $\eta_m\in\mathbb{R}^p$, $\|\eta_m\|=1$ denote an arbitrary normalized eigenvector of M_m with corresponding eigenvalue λ_m, $m\geq 1$. Then,

(a) $\lim_{m\to\infty}\lambda_m=0$.
(b) If the vectors $\{Wu_i\}_{i=1}^{k}$ are linearly independent, then $\lim_{m\to\infty}\eta_m^{\top}v_i=0$ for $i=1,\ldots,k$.

Proof. To begin, we shall establish part (a). Let $M=W\sum_{i=1}^{k}u_i v_i^{\top}$. We first note that all p eigenvalues of M must be zero. To see this, let λ denote an arbitrary real eigenvalue of M with corresponding eigenvector v (possibly complex) such that

$\bar{v}^\top v = 1$, where $\bar{v}$ denotes the complex conjugate of v. Because $M^2 = 0_{p \times p}$ from $v_i^\top W = 0_p^\top$ for $i = 1, \ldots, k$ and $Mv = \lambda v$, we have

$$0 = \bar{v}^\top 0_p = \bar{v}^\top 0_{p \times p} v = \bar{v}^\top M^2 v = \bar{v}^\top M(Mv) = \bar{v}^\top (Mv)\lambda = \bar{v}^\top v \lambda^2 = \lambda^2,$$

implying $\lambda = 0$.

Writing $\lambda_{m,1}, \ldots, \lambda_{m,p}$ to denote the p real eigenvalues of M_m, the Ostrowski-Elsner Theorem [cf. Stewart and Sun (1990), p. 170] states that we can bound the maximum difference between the eigenvalues $\{\lambda_{m,i}\}_{i=1}^p$ of M_m and the eigenvalues (all zero) of the limiting matrix M as

$$\max_{1 \leq i \leq p} |\lambda_{m,i}| = \max_{1 \leq i \leq p} |\lambda_{m,i} - 0| \leq (2p - 1) \left(\|M\|_{\mathbb{F}} + \|M_m\|_{\mathbb{F}} \right)^{1 - \frac{1}{p}} \|M_m - M\|_{\mathbb{F}}^{\frac{1}{p}}.$$

Because $\lim_{m \to \infty} M_m = M$ and the Frobenius matrix norm is continuous, we have from the above inequality that $\lim_{m \to \infty} \max_{1 \leq i \leq p} |\lambda_{m,i}| = 0$. This establishes part (a).

To show part (b), it suffices to show that, for any subsequence $\{m_j\}_{j=1}^\infty$ of $\{m\}_{m=1}^\infty$, there exists a further subsequence $\{m_{j_n}\}_{n=1}^\infty$ of $\{m_j\}_{j=1}^\infty$ such that $\lim_{n \to \infty} v_i^\top \eta_{m_{j_n}} = 0$ for each $i = 1, \ldots, k$. By the Bolzano-Weierstrass theorem, there exists a convergent subsequence $\{m_{j_n}\}_{n=1}^\infty$ of $\{m_j\}_{j=1}^\infty$ such that $\lim_{n \to \infty} \eta_{m_{j_n}} = c \in \mathbb{R}^p$, $\|c\| = 1$.

Note $\|\eta_m\| = 1$ is bounded and $M_m \eta_m = \lambda_m \eta_m$, for all $m \geq 1$. As $m \to \infty$, we find that

$$M \eta_m = -(M_m - M)\eta_m + \lambda_m \eta_m \to 0_p$$

follows from $\lim_{m \to \infty} \lambda_m = 0$ and $\lim_{m \to \infty}(M_m - M) = 0_{p \times p}$. From the above limit of $M \eta_m$, it stands that

$$0_p = \lim_{n \to \infty} M \eta_{m_{j_n}} = \sum_{i=1}^k a_i \cdot W u_i,$$

where $a_i = v_i^\top c = \lim_{n \to \infty} v_i^\top \eta_{m_{j_n}}$, $i = 1, \ldots, k$. Because $\{W u_i\}_{i=1}^k$ are linearly independent, it must be the case that each $a_i = 0$ or, equivalently, $\lim_{n \to \infty} \eta_{m_{j_n}}^\top v_i = 0$ for $i = 1, \ldots, k$. This completes the proof of part (b). $\square$

Data contamination

To establish a bound on the finite sample breakdown point of SIR in the case of an unknown covariance structure, we use k-slice (replacement) contamination scheme as in Section 4.3.1 with the following exception. Instead of (4.22), the contamination of the x-values is here assumed to be of the following structure: $\widetilde{x}_{1_h,m} = t_m\widetilde{\beta}_h + \widetilde{v}_h$, $\widetilde{\beta}_h, \widetilde{v}_h \in \mathbb{R}^p$ for $h = 1, \ldots, k$, where

$$\{\widetilde{\beta}_h\}_{h=1}^k \text{ are linearly independent.} \tag{4.39}$$

This provides a contaminated sample $(X, Y)_m^{n,k}$ based on replacing the first observation in the first k slices.

To estimate the unknown covariance Σ using the data $(X, Y)_m^{n,k}$, we compute the sample covariance matrix $\widehat{\Sigma}_m$ of $(X, Y)_m^{n,k}$ which can be algebraically rewritten as

$$\begin{aligned}
\widehat{\Sigma}_m &= \frac{n-k}{n}\widehat{\Sigma}^* + \frac{n-k}{n^2}\sum_{h=1}^k (\widetilde{x}_{1_h,m} - \overline{x}^*)(\widetilde{x}_{1_h,m} - \overline{x}^*)^\top \\
&\quad + \frac{k}{n^2}\sum_{h=1}^k (\widetilde{x}_{1_h,m} - \overline{x}_{k,m})(\widetilde{x}_{1_h,m} - \overline{x}_{k,m})^\top,
\end{aligned}$$

for

$$\widehat{\Sigma}^* = \frac{1}{n-k}\sum_{\substack{i\neq 1_h \\ h\in\{1,\ldots,k\}}} (x_i - \overline{x}^*)(x_i - \overline{x}^*)^\top, \quad \overline{x}^* = \frac{1}{n-k}\sum_{\substack{i\neq 1_h \\ h\in\{1,\ldots,k\}}} x_i, \quad \overline{x}_{k,m} = \frac{1}{k}\sum_{h=1}^k \widetilde{x}_{1_h,m}.$$

For our purposes, it is helpful to further rewrite $\widehat{\Sigma}_m$ as

$$\widehat{\Sigma}_m = S + \frac{n-k}{n^2}t_m^2\sum_{h=1}^k \beta_{h,m}\beta_{h,m}^\top + \frac{k}{n^2}t_m^2\sum_{h=1}^k (\beta_{h,m} - \overline{\beta}_m)(\beta_{h,m} - \overline{\beta}_m)^\top \tag{4.40}$$

where

$$S = \frac{n-k}{n}\widehat{\Sigma}^*, \quad \overline{\beta}_m = \frac{1}{k}\sum_{h=1}^k \beta_{h,m}, \quad \beta_{h,m} = \widetilde{\beta}_h + \frac{\widetilde{v}_h - \overline{x}^*}{t_m}, \; h = 1, \ldots, k.$$

Because we determine the dimension reduction space by multiplying $\mathcal{K}$ normalized eigenvectors of (4.28) by $\widehat{\Sigma}^{-1/2}$, or equivalently by finding $\mathcal{K}$ eigenvectors of $\widetilde{V}$ from

(4.29), it is important to understand how these matrices behave in the presence of data contamination. The next result addresses this point.

In the following, let $\widetilde{\mathcal{B}} \subset \mathbb{R}^p$ denote the space spanned by $\{\widetilde{\beta}_h\}_{h=1}^k$ and write $\overline{x}_h^* = \sum_{i \neq 1_h : y_i \in \mathbb{I}_h} x_i / (n_h - 1)$ to denote the sample mean of the uncontaminated observations in the hth slice, $h = 1, \ldots, k$.

Theorem 4.2 *Under the k-slice contamination scheme described in (4.39) (see also Section 4.3.1), $1 \leq k \leq \min\{p, H\}$, suppose a contaminated sample $(X, Y)_m^{n,k}$ yields an estimate $\widehat{\Sigma}_m$ of Σ as well as a $p \times p$ matrix $\widetilde{V}_m$ from (4.29). Assuming the matrix S is positive definite in (4.40), it holds that:*

$$\lim_{m \to \infty} \widehat{\Sigma}_m^{-1} = S^{-1/2} \left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2}, \tag{4.41}$$

$$\lim_{m \to \infty} \frac{1}{t_m} \widetilde{V}_m = S^{-1/2} \left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2} \sum_{h=1}^k \frac{n_h - 1}{n_h n} (\overline{x}_h^* - \overline{x}^*) \widetilde{\beta}_h^\top. \tag{4.42}$$

Remark. For the case that the covariance matrix Σ of the regressor variables X is unknown, the above Theorem has powerful implications as to how k-slice contamination will influence the e.d.r. subspace resulting from the SIR procedure under contamination. The behavior of the inverse of the covariance matrix $\widehat{\Sigma}_m^{-1}$, estimated from contaminated data $(X, Y)_m^{n,k}$, is key because this matrix appears in the contaminated matrix version $\widetilde{V}_m$ of (4.29), where $\widetilde{V}_m$ in turn is used for obtaining the contaminated e.d.r directions as its eigenvectors. In the limiting form of contamination (i.e., as $t_m \to \infty$), the positive definite inverse matrix $\widehat{\Sigma}_m^{-1}$ actually degenerates to a singular matrix in (4.41). But more importantly, the null space of $\lim_{m \to \infty} \widehat{\Sigma}_m^{-1}$ is *precisely* the space spanned by the k linearly independent directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^k$ from (4.39). That is, the orthogonal complement of the column space of $\lim_{m \to \infty} \widehat{\Sigma}_m^{-1} = S^{-1/2} \left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2}$ is exactly the $\mathbb{R}^p$-space $\mathrm{span}(\{\widetilde{\beta}_h\}_{h=1}^k)$. The significance of this result is the following. Suppose $\widehat{\beta}_m$ represents an estimated e.d.r. direction arising from the SIR procedure applied to the contaminated data $(X, Y)_m^{n,k}$. The vector $\widehat{\beta}_m$ is an eigenvector of $\widetilde{V}_m$, i.e.,

$$\widehat{\Sigma}_m^{-1} c_m = \widetilde{V}_m \widehat{\beta}_m = \lambda_m \widehat{\beta}_m,$$

where $c_m \in \mathbb{R}^p$ is given by

$$c_m = \widehat{\Sigma}_m^{-1} \left(\sum_{h=1}^{k} \widehat{p}_h (\overline{x}_{h,m} - \overline{x}_m)(\overline{x}_{h,m} - \overline{x}_m)^\top + \sum_{h=k+1}^{H} \widehat{p}_h (\overline{x}_h - \overline{x}_m)(\overline{x}_h - \overline{x}_m)^\top \right) \widehat{\beta}_m,$$

involving the contaminated sample means from the first k slices $\{\overline{x}_{h,m}\}_{h=1}^{k}$ and the contaminated overall sample mean $\overline{x}_m$. Hence, we expect the e.d.r. direction $\widehat{\beta}_m$ to be in the column space of $\widehat{\Sigma}_m^{-1}$ so that $\lim_{m\to\infty} \widehat{\beta}_m$ should be in the column space of $\lim_{m\to\infty} \widehat{\Sigma}_m^{-1}$ and therefore *orthogonal* to the directions of contamination $\{\widetilde{\beta}_h\}_{h=1}^{k}$ from (4.39). In fact, the limit (as $m \to \infty$) of any eigenvector of $\widetilde{V}_m$, or equivalently $\widetilde{V}_m/t_m$, can be expected to 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 (4.42) and hence orthogonal to $\mathrm{span}(\{\widetilde{\beta}_h\}_{h=1}^{k})$.

Proof. The result in (4.41) follows immediately from applying Lemma 4.2(b) to the form of $\widehat{\Sigma}_m$ in (4.40).

To establish (4.42), we first note that the matrix $\widetilde{V}$ from (4.29) can be generally rewritten in a form more suitable for our purposes. With some algebra, it holds that

$$\sum_{i=1}^{n}(x_i-\overline{x})(x_i-\overline{x})^\top = \sum_{h=1}^{H} S_h + \sum_{h=1}^{H} n_h(\overline{x}_h-\overline{x})(\overline{x}_h-\overline{x})^\top, \quad S_h = \sum_{i:y_i\in\mathbb{I}_h}(x_i-\overline{x}_h)(x_i-\overline{x}_h)^\top,$$

so that upon division by the sample size n, we have

$$\widehat{\Sigma} = \frac{1}{n}\left(\sum_{h=1}^{H} S_h\right) + \sum_{h=1}^{H} \widehat{p}_h(\overline{x}_h - \overline{x})(\overline{x}_h - \overline{x})^\top,$$

from which it follows that

$$\widetilde{V} = \widehat{\Sigma}^{-1} \sum_{h=1}^{H} \widehat{p}_h(\overline{x}_h - \overline{x})(\overline{x}_h - \overline{x})^\top = \mathrm{I}_p - \widehat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{h=1}^{H} S_h\right).$$

Now consider computing a version $\widetilde{V}_m$ of $\widetilde{V}$ under contamination. Under the k-slice contamination scheme, the sums S_h for $h = k+1, \ldots, H$ remain unchanged and are not functions of m. However, under contamination affecting the first observation $(\widetilde{x}_{1_h,m}, \widetilde{y}_{1_h} = y_{1_h})$ in the first k-slices ($h = 1, \ldots, k$), we may write contaminated versions, say $\{S_{h,m}\}_{h=1}^{k}$, of the first k sums $\{S_h\}_{h=1}^{k}$ as a function of $\{\widetilde{x}_{1_h,m}\}_{h=1}^{k}$ by

$$S_{h,m} = S_h^* + \frac{n_h - 1}{n_h} t_m^2 \left(\beta_{h,m} + \frac{\overline{x}^* - \overline{x}_h^*}{t_m} \right) \left(\beta_{h,m} + \frac{\overline{x}^* - \overline{x}_h^*}{t_m} \right)^\top, \quad h = 1, \ldots, k,$$

with $S_h^* = \sum_{i \neq 1_h : y_i \in \mathbb{I}_h} (x_i - \overline{x}_h^*)(x_i - \overline{x}_h^*)^\top$ for $h = 1, \ldots, k$, and $\beta_{h,m}$ as defined in (4.40). Hence, writing

$$E_m = \mathrm{I}_p - \frac{1}{n} \widehat{\Sigma}_m^{-1} \left(\sum_{h=1}^{k} S_h^* + \sum_{h=k+1}^{H} S_h \right), \qquad u_h = \overline{x}^* - \overline{x}_h^*, \; h = 1, \ldots, k,$$

we can compute the contaminated $\widetilde{V}_m$ as

$$
\begin{aligned}
\widetilde{V}_m &= \mathrm{I}_p - \widehat{\Sigma}_m^{-1} \left(\frac{1}{n} \sum_{h=1}^{k} S_{h,m} + \frac{1}{n} \sum_{h=k+1}^{H} S_h \right) \\
&= E_m - \widehat{\Sigma}_m^{-1} \sum_{h=1}^{k} \frac{n_h - 1}{n_h n} t_m^2 \left(\beta_{h,m} + \frac{\overline{x}^* - \overline{x}_h^*}{t_m} \right) \left(\beta_{h,m} + \frac{\overline{x}^* - \overline{x}_h^*}{t_m} \right)^\top \\
&= E_m - \sum_{h=1}^{k} \frac{n_h - 1}{n_h n} \left(t_m^2 \widehat{\Sigma}_m^{-1} \beta_{h,m} + t_m \widehat{\Sigma}_m^{-1} u_h \right) \left(\beta_{h,m} + \frac{u_h}{t_m} \right)^\top .
\end{aligned}
$$

Upon division by m, we have

$$\frac{1}{t_m} \widetilde{V}_m = \frac{1}{t_m} E_m - \sum_{h=1}^{k} \frac{n_h - 1}{n_h n} \left(t_m \widehat{\Sigma}_m^{-1} \beta_{h,m} + \widehat{\Sigma}_m^{-1} u_h \right) \left(\beta_{h,m} + \frac{u_h}{t_m} \right)^\top .$$

Applying Lemma 4.2 with the form of $\widehat{\Sigma}_m$ in (4.40), we have that

$$\lim_{m \to \infty} \frac{1}{t_m} \widehat{\Sigma}_m^{-1} = 0_{p \times p}, \qquad \lim_{m \to \infty} t_m \widehat{\Sigma}_m^{-1} \beta_{h,m} = 0_p, \qquad \lim_{m \to \infty} u_h \left(\beta_{h,m} + \frac{u_h}{t_m} \right)^\top = u_h \widetilde{\beta}_h^\top,$$

so that $\lim_{m \to \infty} E_m / t_m = 0_{p \times p}$ and the form of $\lim_{m \to \infty} \widetilde{V}_m / t_m$ follows from using the limit of $\widehat{\Sigma}_m^{-1}$. $\square$

In accordance with Definition 4.1, we can now derive an upper bound on the finite sample breakdown point of a SIR-type procedure with an unknown covariance structure and known dimension $\mathcal{K}$ of reduction. To derive the bound, we will require the following mild definition of a *breakdown pattern* with respect to the data $(X, Y)^n$. The breakdown pattern implies that the constitution of the data $(X, Y)^n$ allows for k observations to be replaced in order to a sequence of contaminated data sets $(X, Y)_m^{n,k}$, $m = 1, 2, \ldots$ which has some mild properties.

For all practical purposes and to facilitate the proofs, we may consider again a k-replacement contamination scheme that involves corrupting the first observation in

each of the first $k \leq H$ slices in the $(X, Y)^n$. However, to allow for more generality in applying the breakdown result, the breakdown pattern definition shall detail conditions for which a more general k-replacement contamination scheme $(X, Y)_m^{n,k}$ will ensure *breakdown* in the sense that the maximal distance $\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k})$ between the uncorrupted $\widehat{\mathcal{B}}$ and corrupted $\widehat{\mathcal{B}}_{m,k}$ subspace estimates, based on $(X, Y)^n$ and $(X, Y)_m^{n,k}$, can be obtained as $m \to \infty$.

Definition 4.2 (*Breakdown pattern for given $\mathcal{K}$ and unknown covariance.*)
Suppose that a size n data sample $(X, Y)^n$ yields a $\mathcal{K}$-dimensional subspace of $\mathbb{R}^p$ based on Sir*-estimated e.d.r. directions $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$, where $1 \leq \mathcal{K} < p$. Let $k = \min\{\mathcal{K}, p - \mathcal{K}\}$. The sample $(X, Y)^n$ is said to satisfy a **breakdown pattern** for given $\mathcal{K}$ and unknown $\mathrm{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$ if, there exists*

1. *some arbitrary selection of k slices from the available H slices, denoted by indices $\{h_j\}_{j=1}^k \subset \{1, \ldots, H\}$;*

2. *some arbitrary selection of k observations from $(X, Y)^n$ involving a single observation, denoted $(x_{i_{h_j}}, y_{i_{h_j}})$, chosen from each of the above k selected slices with indices $\{h_j\}_{j=1}^k$: i.e., $y_{i_{h_j}} \in \mathbb{I}_{h_j}$ for each $j = 1, \ldots, k$;*

3. *some arbitrary selection of k e.d.r. directions, denoted $\{\widehat{\beta}_{i_j}\}_{j=1}^k$, from the e.d.r. directions $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$ estimated from $(X, Y)^n$;*

4. *a positive real sequence $\{t_m\}_{m=1}^{\infty}$ such that $t_m \to \infty$ as $m \to \infty$;*

such that:

(i) for any real constants $\{a_j\}_{j=1}^k$,

$$\sum_{j=1}^k a_j \cdot \frac{n_{h_j} - 1}{n_{h_j}} (\overline{x}_{-h_j}^* - \overline{x}_-^*) \in \mathrm{span}\{\widehat{\beta}_{i_j}\}_{j=1}^k \text{ only if } a_1 = \cdots = a_k = 0,$$

and

$$S_- = \frac{1}{n} \sum_{\substack{i=1 \\ i \notin \{i_{h_1}, \dots, i_{h_k}\}}}^{n} (x_i - \overline{x}^*_-)(x_i - \overline{x}^*_-)^\top \ \text{is positive definite,}$$

where

$$\overline{x}^*_{-h_j} = \frac{1}{n_{h_j} - 1} \sum_{\substack{i:y_i \in \mathbb{I}_{h_j} \\ i \neq i_{h_j}}} x_i, \ j = 1, \dots, k; \qquad \overline{x}^*_- = \frac{1}{n-k} \sum_{\substack{i=1 \\ i \notin \{i_{h_1}, \dots, i_{h_k}\}}}^{n} x_i,$$

denote the sample mean $\overline{x}^*_{-h_j}$ of the h_jth slice after deleting the selected observation $x_{i_{h_j}}$ $(j = 1, \dots, k)$ along with the overall sample mean $\overline{x}^*_-$ and a covariance estimate S_- of Σ after deleting the k selected observations $\{x_{i_{h_j}}\}_{j=1}^k$ from $(X, Y)^n$;

(ii) for some size k subset $\widehat{\beta}_{m,\ell_1}, \dots, \widehat{\beta}_{m,\ell_k}$ of the first $\mathcal{K}$ SIR-estimated e.d.r. directions from the contaminated data $(X, Y)_m^{n,k}$, found by replacing $(x_{i_{h_j}}, y_{i_{h_j}}) \in (X, Y)^n$ with $(\widetilde{x}_{i_{h_j}}, y_{i_{h_j}})$ where

$$\widetilde{x}_{i_{h_j}} = t_m \beta_{i_j} + \widetilde{v}_j, \qquad \widetilde{v}_j \in \mathbb{R}^p, \ j = 1, \dots, k, \tag{4.43}$$

then the $p \times k$ matrix $B_m = [\widehat{\beta}_{m,\ell_1} \cdots \widehat{\beta}_{m,\ell_k}]$ satisfies

$$\liminf_{m \to \infty} \frac{\det(B_m^\top B_m)}{\prod_{j=1}^k \|\widehat{\beta}_{j,m}\|^2} \geq C, \tag{4.44}$$

for some $C > 0$.

Note: *If the first k slices are selected (i.e., $h_j = j$ for $j = 1, \dots, k$) and the first observation in each slice is selected (i.e., $x_{i_{h_j}} = x_{1_j}$ for $j = 1, \dots, k$) to satisfy (i)-(ii) in the above definition, then the k-replacement contamination scheme (4.43) is the same as (4.39) and $\overline{x}^*_{-h_j} = \overline{x}^*_j$ for $j = 1, \dots, k$, $\overline{x}^*_- = \overline{x}^*$, $S_- = S$, where $\overline{x}^*_j$, $\overline{x}^*$, S correspond to the quantities appearing in (4.42) of Theorem 4.2.*

Using the first k slice replacement scheme (4.39), the condition (i) in Definition 4.2 may be expressed as: for any real constants $\{a_j\}_{j=1}^k$,

$$\sum_{h=1}^k a_h \cdot \frac{n_h - 1}{n_h}(\overline{x}^*_h - \overline{x}^*) \in \text{span}\{\widehat{\beta}_{i_j}\}_{j=1}^k \quad \text{only if} \quad a_1 = \dots = a_k = 0,$$
$$\tag{4.45}$$

and the $p \times p$ matrix S from (4.40) is positive definite.

For the $\mathcal{K} = 1$ case involving contamination of the first observation in the first slice, Definition 4.2 (specifically, (4.45)) is an alternative, but equivalent, formulation of the conditions required in Gather et al. (2002) and Hilker (1997) for studying errors in direction estimation with unknown covariance structure. In addition, (4.44) is a weaker version of a similar condition used in Hilker (1997, Corollary 4.20). Both (4.45) and (4.44) are technical conditions that facilitate the proof for the breakdown point, given $\mathcal{K}$ and unknown covariance $\mathrm{Cov}(X) = \Sigma$. Again the main point of the contamination used here is that regressor observations are contaminated by corrupting observations located in different slices (4.43), but for discussion purposes it suffices to consider contamination $(X, Y)_m^{n,k}$ using the first observation in the first k slices as in (4.39).

We now give a bound on the breakdown point.

Corollary 4.2 *For a given $1 \leq \mathcal{K} < p$, suppose the SIR procedure seeks to estimate a $\mathcal{K}$-dimensional subspace of $\mathbb{R}^p$ based on $H \geq \min\{\mathcal{K}, p - \mathcal{K}\}$ data slices with a size n data sample $(X, Y)^n$ and unknown values of $\mathrm{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$. Provided that the data $(X, Y)^n$ satisfies the Breakdown Pattern Definition 4.2,*
(i) there exists a sequence $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, of contaminated data sets and associated subspace estimates $\widehat{\mathcal{B}}_{m,k}$, found by replacing $k = \min\{\mathcal{K}, p - \mathcal{K}\}$ observations in $(X, Y)^n$, where

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

where $\widehat{\mathcal{B}}$ is the subspace estimate based on $(X, Y)^n$.
(ii) the finite sample breakdown point of SIR, under Definition 4.1, satisfies

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

Remark. We again make a few comments on the nature of the proof and the contamination scheme used.

The dimension $\mathcal{K}$ of the reduction subspace is known and the uncontaminated data $(X, Y)^n$ produces a $\mathcal{K}-$dimensional estimate $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ of the reduction subspace, while

$\widehat{\mathcal{B}}_{m,k} \subset \mathbb{R}^p$ represents a corrupted $\mathcal{K}$−dimensional estimate of the reduction subspace, based on replacing k observations in $(X,Y)^n$. WloG and to ease our discussion, we can assume that we shall contaminate the first observation in each of the first $\mathcal{K}$ slices as in the contamination plan of (4.39). The maximal distance $\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k})$ between the corrupted and uncorrupted subspace estimates is again given in (4.1). There are the two possible cases for this distance, depending on $2\mathcal{K} \leq p$ or $2\mathcal{K} > p$.

If $2\mathcal{K} \leq p$, the subspace estimates $\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k}$ are maximally distant under $\mathbb{F}$ if these estimates are orthogonal. If the uncontaminated data $(X,Y)^n$ are in the breakdown pattern, we replace $k = \mathcal{K}$ observations with contaminated vectors as in (4.39), where the directions of contamination correspond to $\widehat{\beta}_1, \ldots, \widehat{\beta}_\mathcal{K}$, namely the SIR-estimated e.d.r. directions from $(X,Y)^n$. Using Theorem 4.2, we show that this approach shall yield a contaminated estimate $\widehat{\mathcal{B}}_{m,k=\mathcal{K}}$ of the e.d.r. subspace that is arbitrary close to being orthogonal to the uncontaminated estimate $\widehat{\mathcal{B}}$ (in the sense that $P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} \approx 0_{p \times p}$).

If $2\mathcal{K} > p$, the subspace estimates $\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,k}$ must have a non-empty intersection that is at least of dimension $2\mathcal{K} - p$ by Lemma 2.2. In this case, subspaces $\widehat{\mathcal{B}}$ and $\widehat{\mathcal{B}}_{m,k}$ are maximally distant under $\mathbb{F}$ if the intersection $\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k}$ is as minimal as possible (i.e., $\dim(\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k}) = 2\mathcal{K} - p$) and, after removing the intersection, the remaining parts of both subspaces $\widehat{\mathcal{B}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp$ and $\widehat{\mathcal{B}}_{m,k} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp$ are orthogonal. Here we can replace $k = p - \mathcal{K}$ observations to obtain a contaminated sample and a contaminated e.d.r. subspace $\widehat{\mathcal{B}}_{m,k=p-\mathcal{K}}$ by choosing $k = p - \mathcal{K} < \mathcal{K}$ directions from among the $\mathcal{K}$ uncontaminated estimates $\widehat{\beta}_1, \ldots, \widehat{\beta}_\mathcal{K}$ to serve as the directions of contamination under the scheme (4.39). Using Theorem 4.2 again with this contamination plan, we can show $\dim(\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,p-\mathcal{K}}) \approx 2\mathcal{K} - p$ and $P_{\widehat{\mathcal{B}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}} \cap (\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}_{m,k})^\perp} \approx 0_{p \times p}$ in the limit (as $m \to \infty$). Hence, this contamination approach provides enough contamination to force breakdown of the SIR dimension reduction procedure.

Proof. Let $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ denote the $\mathcal{K}$-dimensional reduction subspace determined from the uncontaminated data; $\widehat{\mathcal{B}} \subset \mathbb{R}^p$ is spanned by the scaled eigenvectors $\widehat{\beta}_1 = \Sigma^{-1/2} \widehat{\eta}_1, \ldots, \widehat{\beta}_\mathcal{K} = \Sigma^{-1/2} \widehat{\eta}_\mathcal{K}$ associated with the $\mathcal{K}$ largest eigenvalues of (4.20) computed with $(X,Y)^n$. To establish a bound on $\epsilon_{fsbp,\mathcal{K}}$, we consider two possible cases:

$2\mathcal{K} \leq p$ or $2\mathcal{K} > p$. For each case, we examine the finite sample breakdown point under different contamination schemes.

CASE I: $2\mathcal{K} \leq p$. Because the data $(X, Y)^n$ satisfy the breakdown pattern Definition 4.2 by assumption, we can get a contaminated sequence $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, as in (4.43) by picking $k = \min\{\mathcal{K}, p - \mathcal{K}\} = \mathcal{K}$ slices from among the available H slices, one observation from each of these $\mathcal{K}$ slices and all $\mathcal{K}$ e.d.r. directions $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$. WloG we assume that the contamination scheme (4.39), with $k = \mathcal{K}$ contamination directions $\widetilde{\beta}_h = \widehat{\beta}_h$, $h = 1, \ldots, \mathcal{K}$, can be used to obtain $(X, Y)_m^{n,k}$ under Definition 4.2; that is, we assume that the first $k = \mathcal{K}$ slices and the first observation in each slice $\{x_{1_h}\}_{h=1}^{\mathcal{K}}$ may be used so that (4.44) and (4.45) hold; if this choice does not satisfy Definition 4.2, then other slices and observations may be used for contamination under (4.43) and the arguments require only notational modifications.

The contaminated data yields a subspace estimate $\widehat{\mathcal{B}}_{m,\mathcal{K}}$ based on eigenvectors $\widehat{\beta}_{1,m}, \ldots, \widehat{\beta}_{\mathcal{K},m}$ corresponding to the largest $\mathcal{K}$ eigenvalues of the contaminated matrix $\widetilde{V}_m$ (computed from using $(X, Y)_m^{n,\mathcal{K}}$ in (4.29)).

Under the first $k = \mathcal{K}$ replacement scheme (4.39), we now apply the results in Lemma 4.3(b) and Theorem 4.2 using conditions (4.44) and (4.45) (since again we assume the data $(X, Y)^n$ are in a breakdown pattern involving the first observation from the first $k = \mathcal{K}$ slices). We may first assume that the matrix S from (4.40) is positive definite under condition (4.45). Consequently, we may apply (4.42) in Theorem 4.2 to find

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

where $\widetilde{\mathcal{B}}$ denotes the $\mathbb{R}^p$-subspace spanned by the directions of contamination $\{\widetilde{\beta}_h = \widehat{\beta}_h\}_{h=1}^{\mathcal{K}}$.

We now wish to apply Lemma 4.3(b) treating $k = \mathcal{K}$, $M_m = \widetilde{V}_m/t_m$, $W = S^{-1/2}\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}}\right) S^{-1/2}$, $u_h = \frac{n_h - 1}{n_h n}(\overline{x}_h^* - \overline{x}^*)$ and $v_h = \widetilde{\beta}_h = \widehat{\beta}_h$ $(h = 1, \ldots, \mathcal{K})$ in the notation of Lemma 4.3. With this notation, it holds that $W^{\top}v_h = Wv_h = 0_p$

for each $h = 1, \ldots, \mathcal{K}$ since $v_h = \widetilde{\beta}_h \in \widetilde{\mathcal{B}}$ so that $P_{S^{-1/2}\widetilde{\mathcal{B}}} S^{-1/2} v_h = S^{-1/2} v_h$. To apply Lemma 4.3(b) to $M_m = \widetilde{V}_m / t_m$, we need to show that the collection $\{W u_h\}_{h=1}^{\mathcal{K}}$ of vectors are linearly independent. Suppose, for some real constants $a_1, \ldots, a_{\mathcal{K}}$, it holds that $\sum_{h=1}^{\mathcal{K}} a_h W u_h = W \sum_{h=1}^{\mathcal{K}} a_h u_h = 0_p$, implying that $(\sum_{h=1}^{\mathcal{K}} a_h u_h)^\top W (\sum_{h=1}^{\mathcal{K}} a_h u_h) = 0$ or equivalently that

$$\left(\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2} \left(\sum_{h=1}^{\mathcal{K}} a_h u_h \right) \right)^\top \left(\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2} \left(\sum_{h=1}^{\mathcal{K}} a_h u_h \right) \right) = 0;$$

this in turn implies that

$$\left(\mathrm{I}_p - P_{S^{-1/2}\widetilde{\mathcal{B}}} \right) S^{-1/2} \left(\sum_{h=1}^{\mathcal{K}} a_h u_h \right) = 0_p \qquad \Rightarrow \qquad S^{-1/2} \left(\sum_{h=1}^{\mathcal{K}} a_h u_h \right) \in S^{-1/2}\widetilde{\mathcal{B}},$$

so that

$$\sum_{h=1}^{\mathcal{K}} a_h u_h \in \widetilde{\mathcal{B}} = \mathrm{span}\{\widetilde{\beta}_h\}_{h=1}^{\mathcal{K}} = \mathrm{span}\{\widehat{\beta}_h\}_{h=1}^{\mathcal{K}}.$$

But because $u_h = \frac{n_h - 1}{n_h n} (\overline{x}_h^* - \overline{x}^*)$ for $h = 1, \ldots, \mathcal{K}$, the above statement can only hold if $a_1 = \cdots = a_{\mathcal{K}} = 0$ by condition (4.45). Hence, the vectors $\{W u_h\}_{h=1}^{\mathcal{K}}$ are linearly independent.

Now note that each vector in $\{\widehat{\beta}_{h,m}^* = \widehat{\beta}_{h,m} / \|\widehat{\beta}_{h,m}\|\}_{h=1}^{\mathcal{K}}$ is an eigenvector of $\widetilde{V}_m / t_m$ in (4.42) and is normalized so that $\|\widehat{\beta}_{h,m}^*\| = 1$ for each $h = 1, \ldots, \mathcal{K}$. Hence, applying Lemma 4.3(b) to $M_m = \widetilde{V}_m / t_m$, we have that for each eigenvector in $\{\widehat{\beta}_{h,m}^*\}_{h=1}^{\mathcal{K}}$ of $\widetilde{V}_m / t_m$ that

$$\lim_{m \to \infty} \widetilde{\beta}_j^\top \widehat{\beta}_{h,m}^* = \lim_{m \to \infty} \widehat{\beta}_j^\top \widehat{\beta}_{h,m}^* = 0 \qquad h = 1, \ldots, \mathcal{K}; j = 1, \ldots, \mathcal{K}. \tag{4.46}$$

Define a $p \times \mathcal{K}$ matrix $B_m^* = [\widehat{\beta}_{1,m}^* \ \widehat{\beta}_{2,m}^* \cdots \widehat{\beta}_{\mathcal{K},m}^*]$ so that the column space of B_m^* is contaminated e.d.r. subspace $\widehat{\mathcal{B}}_{m,\mathcal{K}}$ and $P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} = B_m^* (B_m^{*\top} B_m^*)^{-1} B_m^{*\top}$. Let $B = [\widehat{\beta}_1 \ \widehat{\beta}_2 \cdots \widehat{\beta}_{\mathcal{K}}]$ be a $p \times \mathcal{K}$ matrix with columns defined by the $\mathcal{K}$ uncontaminated e.d.r. directions so that $P_{\widehat{\mathcal{B}}} = B(B^\top B)^{-1} B^\top$.

By (4.44), there exists some $C > 0$ (not depending on m) and some $N \in \mathbb{N}$ such that, for $m \geq N$,

$$\det(B_m^{*\top} B_m^*) = \frac{\det(B_m^\top B_m)}{\prod_{j=1}^{\mathcal{K}} \|\widehat{\beta}_{j,m}\|^2} \geq \frac{C}{2},$$

and $\lim_{m\to\infty} B^\top B_m^* = 0_{\mathcal{K}\times\mathcal{K}}$ by (4.46), so that

$$\lim_{m\to\infty} B^\top P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} = 0_{\mathcal{K}\times\mathcal{K}}$$

follows from Lemma 7.5.

Hence, $\lim_{m\to\infty} P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}} = 0_{p\times p}$ so that

$$\lim_{m\to\infty} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,\mathcal{K}}) = \lim_{m\to\infty} \sqrt{2\mathcal{K} - 2\mathrm{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,\mathcal{K}}})} = \sqrt{2\mathcal{K}}$$

by the continuity of matrix trace operations. By Definition 4.1, we conclude

$$\epsilon_{fsbp}((X,Y)^n, \mathbb{F}, \mathcal{K}) \leq \frac{\mathcal{K}}{n} = \frac{\min\{\mathcal{K}, p-\mathcal{K}\}}{n}.$$

CASE II: $2\mathcal{K} > p$. Because the data $(X,Y)^n$ satisfy the breakdown pattern Definition 4.2 by assumption, we can again get a contaminated sequence $(X,Y)_m^{n,k}$, $m \in \mathbb{N}$, as in (4.43) by picking $k = \min\{\mathcal{K}, p-\mathcal{K}\} = p - \mathcal{K}$ slices from among the available H slices, one observation from each of these $p - \mathcal{K}$ slices and selecting $p - \mathcal{K} < \mathcal{K}$ e.d.r. directions, say $\{\widehat{\beta}_{i_j}\}_{j=1}^{p-\mathcal{K}}$, from among the $\mathcal{K}$ uncontaminated $\widehat{\beta}_1, \ldots, \widehat{\beta}_{\mathcal{K}}$. WloG we may assume $\widehat{\beta}_{i_j} = \widehat{\beta}_j$ for $j = 1, \ldots, p-\mathcal{K}$, i.e., the first $p - \mathcal{K}$ uncontaminated e.d.r. directions are chosen. WloG we also assume that the contamination scheme (4.39), with $k = p - \mathcal{K}$ contamination directions $\widetilde{\beta}_h = \widehat{\beta}_h$, $h = 1, \ldots, p - \mathcal{K}$, can be used to obtain $(X,Y)_m^{n,k}$ under Definition 4.2; that is, we assume that the first $k = p - \mathcal{K}$ slices and the first observation in each slice $\{x_{1_h}\}_{h=1}^{p-\mathcal{K}}$ may be used so that (4.44) and (4.45) hold (where (4.45) uses $\widehat{\beta}_{i_j} = \widehat{\beta}_j$ for $j = 1, \ldots, p-\mathcal{K}$).

The contaminated data yields a subspace estimate $\widehat{\mathcal{B}}_{m,p-\mathcal{K}}$ based on eigenvectors $\widehat{\beta}_{m,1}, \ldots, \widehat{\beta}_{m,\mathcal{K}}$ corresponding to the largest $\mathcal{K}$ eigenvalues of the contaminated matrix $\widetilde{V}_m$ (computed from using $(X,Y)_m^{n,p-\mathcal{K}}$ in (4.29)).

Define a $p \times (p - \mathcal{K})$ matrix $B^{(1)} = [\widehat{\beta}_1\, \widehat{\beta}_2 \cdots \widehat{\beta}_{p-\mathcal{K}}]$ using the $p - \mathcal{K}$ directions of contamination (i.e., the first $p - \mathcal{K}$ uncontaminated e.d.r. directions used in (4.45)). With the same essential arguments of Case I by replacing B there with $B^{(1)}$ (i.e., applying Lemma 4.3(b) and Theorem 4.2 using condition (4.45) and then using (4.44) with Lemma 7.5), it can be shown that

$$\lim_{m\to\infty} B^{(1)\top} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} = 0_{(p-\mathcal{K})\times(p-\mathcal{K})}.$$

From this, if $\widehat{\mathcal{B}}^{(1)} \subset \widehat{\mathcal{B}}$ denotes the column space of $B^{(1)}$ with corresponding projection matrix $P_{\widehat{\mathcal{B}}^{(1)}} = B^{(1)}(B^{(1)\top}B^{(1)})^{-1}B^{(1)\top}$, we have

$$\lim_{m\to\infty} P_{\widehat{\mathcal{B}}^{(1)}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}} = 0_{p\times p}. \tag{4.47}$$

The subspace $\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp} \subset \widehat{\mathcal{B}}$, consisting of the $\widehat{\mathcal{B}}$-vectors orthogonal to $\widehat{\mathcal{B}}^{(1)}$, has projection matrix $P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}} = P_{\widehat{\mathcal{B}}} - P_{\widehat{\mathcal{B}}^{(1)}}$. We note that $\mathrm{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) \geq 2\mathcal{K} - p$ by (4.26) and

$$
\begin{aligned}
\mathrm{tr}(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) &\leq \mathrm{rank}(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}}) \\
&= \mathrm{tr}(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}}) \\
&= \mathrm{tr}(P_{\widehat{\mathcal{B}}} - P_{\widehat{\mathcal{B}}^{(1)}}) \\
&= \mathrm{tr}(P_{\widehat{\mathcal{B}}}) - \mathrm{tr}(P_{\widehat{\mathcal{B}}^{(1)}}) \\
&= \mathrm{rank}(P_{\widehat{\mathcal{B}}}) - \mathrm{rank}(P_{\widehat{\mathcal{B}}^{(1)}}) = 2\mathcal{K} - p.
\end{aligned}
$$

Then by (4.47), we have that

$$
\begin{aligned}
2\mathcal{K} - p &\leq \lim_{m\to\infty} \mathrm{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) \\
&\leq \lim_{m\to\infty} \mathrm{tr}\Big((P_{\widehat{\mathcal{B}}^{(1)\perp}} + P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}}) P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}\Big) \\
&\leq \lim_{m\to\infty} \mathrm{tr}\Big(P_{\widehat{\mathcal{B}}^{(1)}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}\Big) + \lim_{m\to\infty} \mathrm{tr}\Big(P_{\widehat{\mathcal{B}} \cap \widehat{\mathcal{B}}^{(1)\perp}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}\Big) \\
&\leq \lim_{m\to\infty} \mathrm{tr}\Big(P_{\widehat{\mathcal{B}}^{(1)}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}\Big) + 2\mathcal{K} - p \\
&\leq 0 + 2\mathcal{K} - p.
\end{aligned}
$$

Hence, $\lim_{m\to\infty} \mathrm{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}}) = 2\mathcal{K} - p$ and so

$$\lim_{m\to\infty} \mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,p-\mathcal{K}}) = \lim_{m\to\infty} \sqrt{2\mathcal{K} - 2\mathrm{tr}(P_{\widehat{\mathcal{B}}} P_{\widehat{\mathcal{B}}_{m,p-\mathcal{K}}})} = \sqrt{2(p-\mathcal{K})}.$$

The limiting value $\sqrt{2(p-\mathcal{K})}$ of $\mathbb{F}(\widehat{\mathcal{B}}, \widehat{\mathcal{B}}_{m,p-\mathcal{K}})$ is the maximal distance between two $\mathcal{K}$-dimensional subspaces of $\mathbb{R}^p$ under $\mathbb{F}$ by (4.1), in the considered case $2\mathcal{K} > p$. By Definition 4.1, we may conclude

$$\epsilon_{fsbp,\mathcal{K}}((X,Y)^n, \mathbb{F}, \mathcal{K}) \leq \frac{p-\mathcal{K}}{n} = \frac{\min\{\mathcal{K}, p-\mathcal{K}\}}{n}. \qquad \square$$

4.4   Sensitivity Analysis when $\mathcal{K}$ is unknown

A careful examination of Sir is required to determine how contamination could possibly influence an estimate $\widehat{\mathcal{K}}$ of $\mathcal{K}$ as well as the estimated subspace $\widehat{B}_{\widehat{\mathcal{K}}}$. Given a data set $(X, Y)^n$, we may replace k data points in (X, Y) to obtain a contaminated data set $(X, Y)^{n,k}_m$. Applying Sir to $(X, Y)^{n,k}_m$ yields a corrupted dimension estimate $\widehat{\mathcal{K}}_k$ and subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k}$. To measure the effect of data contamination on the subspace estimation, we consider the distances between the subspaces $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} \subset \mathbb{R}^p$ (estimated e.d.r. subspace using uncontaminated data $(X, Y)^n$) and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k} \subset \mathbb{R}^p$ (e.d.r. subspace estimated by contaminating k data points $(X, Y)^{n,k}_m$) with the metric $\mathbb{F}$. The worst-case contamination scenario would cause the subspaces $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k}$ to be most distant under the measure $\mathbb{F}$ and we could refer to this occurrence as the *breakdown* of the dimension reduction procedure.

When $\mathcal{K}$ is unknown, we can quantify the breakdown for dimension reduction, or worst-case scenario in contamination with respect to the subspace metric $\mathbb{F}$, as

$$\mathbb{F}(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}, \widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k}) = \sqrt{p} \tag{4.48}$$

using Theorem 2.1. Breakdown occurs when subspace estimates with and without contamination, $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k}$, are maximally distant in terms of the subspace distance measure $\mathbb{F}$. In fact, the corrupted and uncorrupted subspace estimates must be distant to the *largest extent possible* for $\mathbb{R}^p$ subspaces. That is, breakdown in (4.48) requires $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k}$ to be the *orthogonal complement* of $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ in $\mathbb{R}^p$, i.e., $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k} = \widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}$. In contrast, when $\mathcal{K}$ is known, then the maximal distance between subspace estimates (4.1) may be much smaller than (4.48) when $\mathcal{K}$ is unknown. In fact, in the case that $2\mathcal{K} > p$, two dimension $\mathcal{K}$ subspace estimates must have a sizable intersection by Lemma 2.2, so that there are limits to the amount of separation between subspace estimates. This is not the case when $\mathcal{K}$ is unknown and worst case contamination now implies a subspace estimate based on k-replacement must be the orthogonal complement of the subspace estimated by the uncorrupted data.

We are now prepared to define a finite sample breakdown point for SIR-type dimension reduction procedures, in the case that the dimension $\mathcal{K}$ of reduction is unknown. This requires a modification of the finite sample breakdown Definition 4.3 when $\mathcal{K}$ is known, because again the notation of *worst case* contamination has changed.

Definition 4.3 *Finite sample breakdown point in dimension reduction ($\mathcal{K}$ unknown). Let $(X,Y)_m^{n,k}$ denote a contaminated sample found by replacing $1 \leq k \leq n$ data points in a data set $(X,Y)^n = \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^p \times \mathbb{R}$ with arbitrary values $\{(\widetilde{x}_{i_j}, \widetilde{y}_{i_j})\}_{j=1}^k$. Let $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_k} \subset \mathbb{R}^p$ denote reduction subspace estimates, of estimated dimensions $0 \leq \widehat{\mathcal{K}} \leq p$ and $0 \leq \widehat{\mathcal{K}}_k \leq p$, based on a given dimension reduction procedure (e.g., SIR) applied to $(X,Y)^n$ and $(X,Y)_m^{n,k}$, respectively. The **finite sample breakdown point** of the dimension reduction procedure is defined as*

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

under the metric $\mathbb{F}$ for the data constellation $(X,Y)^n$.

The finite sample breakdown point definition in Definition 4.3 quantifies the amount of data in $(X,Y)^n$ that need to be corrupted so that the resulting subspace estimate is completely orthogonal (i.e., the orthogonal complement) of the subspace estimated intended by the original data. Clearly, this type of data contamination entails a dramatic, and arguably the most drastic departure from the subspace estimate of the original data.

Using Definition 4.3, we next examine the robustness of SIR when the dimension of the reduction space must be estimated. In Section 4.4.1, we handle the case that the covariance structure of the regressor variables is known. The robustness of SIR in applications of an unknown covariance structure and unknown dimension $\mathcal{K}$ is detailed in Section 4.4.2.

4.4.1 $\mathcal{K}$ unknown $-\ \Sigma$ known case

To study the robustness of SIR, suppose some data $(X, Y)^n$ produces a subspace esti-
mate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} = \mathrm{span}(\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{\mathcal{K}}})$, based on the first $\widehat{\mathcal{K}}$ e.d.r. directions for some estimate
$\widehat{\mathcal{K}} < p$ of $\mathcal{K}$. Typically, we would expect $\widehat{\mathcal{K}} \geq 1$ to hold, but $\widehat{\mathcal{K}} = 0$ is also allowed in
the following discussion, in which case $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} = 0_p \in \mathbb{R}^p$.

We shall use the k-replacement scheme from Section 4.3.1 to obtain a sequence of
contaminated samples $(X, Y)^{n,k}_m$, $m \in \mathbb{N}$, as follows. Let $k = p - \widehat{\mathcal{K}} \leq H$ and, following
(4.21), replace the first observation (x_{1_h}, y_{1_h}) in each of the first $p - \widehat{\mathcal{K}}$ slices with
$(\widetilde{x}_{1_h}, y_{1_h})$ for

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

where $\{t_m\}_{m=1}^{\infty}$ is a positive real sequence such that $t_m \to \infty$ as $m \to \infty$. Furthermore,
we suppose the $p - \widehat{\mathcal{K}}$ contamination directions have form $\widetilde{\beta}_h = \Sigma^{1/2} \overline{\beta}_h$, $h = 1, \ldots, p - \widehat{\mathcal{K}}$,
where $\{\overline{\beta}_h\}_{h=1}^{p-\widehat{\mathcal{K}}}$ is an orthonormal basis for $(\mathrm{span}(\Sigma^{1/2}\widehat{\beta}_1, \ldots, \Sigma^{1/2}\widehat{\beta}_{\widehat{\mathcal{K}}}))^{\perp}$, the orthogonal
complement of the space $\Sigma^{1/2}\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$. Hence, the vectors $\{\widetilde{\beta}_h\}_{h=1}^{p-\widehat{\mathcal{K}}}$ satisfy (4.22).

We remark that the choice of contaminating the first $p - \widehat{\mathcal{K}}$ is again somewhat
arbitrary and is made here for simplicity. The important idea of contamination is to
contaminate a single observation in each of $k = p - \widehat{\mathcal{K}}$ different slices in a manner
analogous to the above mentioned contamination scheme.

Computing (4.23) (with $k = p - \widehat{\mathcal{K}}$) based on $(X, Y)^{n,k}_m$ yields a contaminated
estimate $\widehat{V}_m$ of $\mathrm{Cov}(\mathrm{E}(\Sigma^{-1/2}X | Y))$ and we shall make the weak assumption that S in
(4.23) (with $k = p - \widehat{\mathcal{K}}$) is positive definite throughout our discussion. The matrix $\widehat{V}_m$
has ordered eigenvalues $\widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ and corresponding orthonormal eigenvectors
$\widehat{\eta}_{m,i}$, $i = 1, \ldots, p$ of $\widehat{V}_m$ and contaminated e.d.r. directions $\widehat{\beta}_{m,i} = \Sigma^{-1/2}\widehat{\eta}_{m,i}$, $i =
1, \ldots, \widehat{\mathcal{K}}$. We now need an estimate $\widehat{\mathcal{K}}_{m,k}$ based on $(X, Y)^{n,k}_m$ to produce a contaminated
subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}} = \mathrm{span}(\widehat{\beta}_{m,1}, \ldots, \widehat{\beta}_{m,\widehat{\mathcal{K}}_{m,k}})$, where $k = p - \widehat{\mathcal{K}}$.

Controlling the estimate $\widehat{\mathcal{K}}_{m,k}$ of the subspace $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ dimension is in general quite
difficult, because this depends on the *size* of the eigenvalues $\widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ as well

as the testing criterion suggested by Li (1991). That is, a determination of the exact magnitude of *all* p contaminated eigenvalues $\{\widehat{\lambda}_{m,i}\}_{i=1}^{p}$ (as a function of m) seems out of reach. At the same time, we give meaning to the sizes of the eigenvalues, and rate their importance, by an arbitrarily chosen significance level used in testing, as pointed out in Ferré (1998). The task of precisely managing the effect of contamination on the estimate $\widehat{\mathcal{K}}_{m,k}$ is then nearly impossible if one does not have the exact critical values needed to test for the dimension of the subspace estimate and in addition one cannot exactly find the eigenvalues to compare against these critical values.

However, we aim to show that the above contamination scheme involving a replacement of $k = p - \widehat{\mathcal{K}}$ observations forces the first $p - \widehat{\mathcal{K}}$ eigenvalues $\{\widehat{\lambda}_{m,p}\}_{h=1}^{p-\widehat{\mathcal{K}}}$ of $\widehat{V}_m$ to become *arbitrarily large* while the last $\widehat{\mathcal{K}}$ eigenvalues $\{\widehat{\lambda}_{m,p}\}_{h=p-\widehat{\mathcal{K}}+1}^{p}$ are bounded and therefore relatively much smaller. This is the most precise result possible that we can expect to achieve in terms of the behavior of the eigenvalues under the contamination.

Note that an explosion of the *only* first $p - \widehat{\mathcal{K}}$ would suggest that the contaminated $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, should lead to a dimension estimate $\widehat{\mathcal{K}}_{m,k}$ where $\widehat{\mathcal{K}}_{m,k} = p - \widehat{\mathcal{K}}$ for large m. If one chooses $\widehat{\mathcal{K}}_{m,k} = p - \widehat{\mathcal{K}}$ as $m \to \infty$ as the contaminated data $(X, Y)_m^{n,k}$ suggests, resulting the contaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ can be shown to be nearly equal to the orthogonal complement of the uncontaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$; namely, for large m, it holds that $P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} \approx 0_{p \times p}$ (i.e., $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ are nearly orthogonal) and $\dim(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}) + \dim(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}) = \widehat{\mathcal{K}} + (p - \widehat{\mathcal{K}}) = p$.

We summarize this finding in the following theorem.

Theorem 4.3 *For a given data set $(X, Y)^n$ with known values of $\mathrm{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$, suppose the SIR procedure yields an estimated reduction subspace $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} \subset \mathbb{R}^p$ with an estimated dimension $0 \leq \widehat{\mathcal{K}} < p$, where $H \leq p$ slices are used in SIR. Then, there exists a sequence $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, of contaminated data sets and associated subspace estimates $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}} \subset \mathbb{R}^p$, found by replacing $k = p - \widehat{\mathcal{K}}$ observations in $(X, Y)^n$, for which the following hold.*

(i) the ordered eigenvalues $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ of $\widehat{V}_m$ from applying the SIR *procedure to $(X,Y)_m^{n,k}$ satisfy*

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

for some $C > 0$;

(ii) there exists $\alpha^* \in (0,1)$ such that if Li (1991)'s testing procedure with significance level $\alpha \in (0, \alpha^*]$ is used to estimate the subspace dimension $\widehat{\mathcal{K}}_{m,k}$, then

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

Remark. To control the Type I error in Li (1991)'s testing procedure for the dimension of the subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$, one should usually choose α to be small in accordance with the condition of Theorem 4.3. This implies that, for many data sets $(X,Y)^n$, it would hold that

$$\epsilon_{fsbp}((X,Y)^n, \mathbb{F}) \leq \frac{p - \widehat{\mathcal{K}}}{n},$$

where $\widehat{\mathcal{K}}$ is the estimated subspace dimension based on $(X,Y)^n$.

Proof. Through the proof, we use the contamination discussed at the beginning of (4.4.1).

We establish part (a) of the theorem first. By (4.23) with $k = p - \widehat{\mathcal{K}}$, we may write

$$\widehat{V}_m = t_m^2 \sum_{h=1}^{p-\widehat{\mathcal{K}}} \frac{\widehat{p}_h}{n_h^2} \left(\overline{\beta}_h + \Sigma^{-1/2} \frac{n_h v_h}{t_m} \right) \left(\overline{\beta}_h + \Sigma^{-1/2} \frac{n_h v_h}{t_m} \right)^\top + S,$$

where we assume S from (4.23) is positive definite so that eigenvalues of $\widehat{V}_m$ are positive, i.e., $0 < \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$.

By part (b) of Theorem 4.1, we have that for the first $p - \widehat{\mathcal{K}}$ eigenvalues $h = 1, \ldots, p - \widehat{\mathcal{K}}$,

$$\lim_{m\to\infty} \frac{\widehat{\lambda}_{m,h}}{t_m^2} = \frac{1}{n_{(p-\widehat{\mathcal{K}}-h+1)}} > 0 \qquad \Rightarrow \qquad \lim_{m\to\infty} \widehat{\lambda}_{m,h} = \infty.$$

As for the last $\widehat{\mathcal{K}}$ eigenvalues $\widehat{\lambda}_{m,p-\widehat{\mathcal{K}}+1}, \ldots, \widehat{\lambda}_{m,p}$, note that by Lemma 4.2, because S is positive definite and the vectors $\{\overline{\beta}_h\}_{h=1}^{p-\widehat{\mathcal{K}}}$ are linearly independent, it holds that

$$\lim_{m\to\infty} \widehat{V}_m^{-1} = S^{-1/2}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}})S^{-1/2},$$

where $\overline{B} = [\overline{\beta}_1 \cdots \overline{\beta}_{p-\widehat{\mathcal{K}}}]$ denotes a $p \times (p - \widehat{\mathcal{K}})$ matrix. Note that

$$
\begin{aligned}
\mathrm{rank}(S^{-1/2}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}})S^{-1/2}) &= \mathrm{rank}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}}) \\
&= \mathrm{tr}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}}) \\
&= \mathrm{tr}(\mathrm{I}_p) - \mathrm{tr}(P_{S^{-1/2}\overline{B}}) \\
&= p - \mathrm{rank}(P_{S^{-1/2}\overline{B}}) = \widehat{\mathcal{K}}.
\end{aligned}
$$

Hence, if $0 \le s_p \le \cdots \le s_1$ denote the ordered eigenvalues of $S^{-1/2}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}})S^{-1/2}$, then $0 < s_{\widehat{\mathcal{K}}} \le \cdots \le s_1$ while $0 = s_p = \cdots = s_{\widehat{\mathcal{K}}+1}$. By Rouché's theorem [Stewart and Sun (1990), p. 167], the ordered eigenvalues $0 < 1/\widehat{\lambda}_{m,1} \le \cdots \le 1/\widehat{\lambda}_{m,p}$ of $\widehat{V}_m^{-1}$ must converge to those of $S^{-1/2}(\mathrm{I}_p - P_{S^{-1/2}\overline{B}})S^{-1/2}$ as $m \to \infty$, namely

$$\lim_{m\to\infty} \frac{1}{\lambda_{m,i}} = s_{p-i+1}, \quad i = 1, \ldots, p,$$

which implies

$$\lim_{m\to\infty} \lambda_{m,h} = \frac{1}{s_{p-h+1}} > 0 \quad \Rightarrow \quad \sup_m |\lambda_{m,h}| \le C_h, \quad h = p - \widehat{\mathcal{K}} + 1, \ldots, p,$$

for some $C_h > 0$, $h = p - \widehat{\mathcal{K}} + 1, \ldots, p$. By setting $C = \max_{p-\widehat{\mathcal{K}}+1 \le h \le p} C_h$, part (a) is now established.

To show part (b), we use the value of C above and find the $\alpha^* \in (0,1)$ such that $\chi^2_{\widehat{\mathcal{K}}(H-p+\widehat{\mathcal{K}}-1),\alpha^*} = nC\widehat{\mathcal{K}}$. This is possible by the continuity of the chi-square distribution. (In the event that $\widehat{\mathcal{K}} = 0$, set $\alpha^* = 1$ and $\chi^2_{0,\alpha^*} \equiv 0$.)

Now we choose a level of significance $\alpha \in (0, \alpha^*]$ in order apply Li (1991)'s test to choose a dimension estimate $\widehat{\mathcal{K}}_{m,k}$ based on the contaminated data $(X,Y)_m^{n,k}$, $m \in \mathbb{N}$ where $k = p - \widehat{\mathcal{K}}$. By the result in part (a), we have that there exists an $N \in \mathbb{N}$ such that $\widehat{\lambda}_{h,m} > \chi^2_{(p-h+1)(H-h),\alpha}$ for $h = 1, \ldots, p - \widehat{\mathcal{K}}$ and $m \ge N$. Hence, for $m \ge N$,

$$t_{h-1} \ge \lambda_{h,m} > \chi^2_{(p-h+1)(H-h),\alpha} \quad h = 1, \ldots, p - \widehat{\mathcal{K}},$$

implying that $p \geq \widehat{\mathcal{K}}_{m,k} \geq p - \widehat{\mathcal{K}}$. By part (a), it holds that for all $m \in \mathbb{N}$,

$$t_{p-\widehat{\mathcal{K}}} = n \sum_{j=p-\widehat{\mathcal{K}}+1}^{p} \widehat{\lambda}_{j,m} \leq nC\widehat{\mathcal{K}} = \chi^2_{\widehat{\mathcal{K}}(H-p+\widehat{\mathcal{K}}-1),\alpha^*},$$

so that we fail to reject the hypothesis $H_0^{p-\widehat{\mathcal{K}}} : \mathcal{K} = p-\widehat{\mathcal{K}}$ and conclude that $\widehat{\mathcal{K}}_{m,k} \leq p-\widehat{\mathcal{K}}$ for all $m \in \mathbb{N}$. Hence, for $m \geq N$, we have $\widehat{\mathcal{K}}_{m,k} = p - \widehat{\mathcal{K}}$.

Therefore, we have immediately that $\lim_{m\to\infty} \widehat{\mathcal{K}}_{m,k} = p - \widehat{\mathcal{K}}$ and also that, for $m \geq N$, $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}} = \operatorname{span}(\widehat{\beta}_{m,1}, \ldots, \widehat{\beta}_{m,\widehat{\mathcal{K}}_{m,k}}) = \operatorname{span}(\widehat{\beta}_{m,1}, \ldots, \widehat{\beta}_{m,k})$, where $k = p - \widehat{\mathcal{K}}$. By Theorem 4.1(e) with $k = p - \widehat{\mathcal{K}}$,

$$\lim_{m\to\infty} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} = \lim_{m\to\infty} P_{\operatorname{span}(\widehat{\beta}_{m,1},\ldots,\widehat{\beta}_{m,k})} = P_{\operatorname{span}(\Sigma^{-1}\widetilde{\beta}_1,\ldots,\Sigma^{-1}\widetilde{\beta}_k)} = P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}},$$

where $\operatorname{span}(\Sigma^{-1}\widetilde{\beta}_1, \ldots, \Sigma^{-1}\widetilde{\beta}_k) = \operatorname{span}(\Sigma^{-1/2}\overline{\beta}_1, \ldots, \Sigma^{-1/2}\overline{\beta}_k) = \widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}$ and $P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}} = 0_{p \times p}$. Finally, we have

$$
\begin{aligned}
\lim_{m\to\infty} \mathbb{F}(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}, \widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}) &= \lim_{m\to\infty} \sqrt{\widehat{\mathcal{K}} + \widehat{\mathcal{K}}_{m,k} - 2\operatorname{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}})} \\
&= \sqrt{\widehat{\mathcal{K}} + (p - \widehat{\mathcal{K}}) - 2\operatorname{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}})} \\
&= \sqrt{p}.
\end{aligned}
$$

This establishes part (b) of the theorem. $\square$

4.4.2 $\mathcal{K}$ unknown $-$ Σ unknown case

Again we suppose some data $(X, Y)^n$ produces a subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} = \operatorname{span}(\widehat{\beta}_1, \ldots, \widehat{\beta}_{\widehat{\mathcal{K}}})$, based on the first $\widehat{\mathcal{K}}$ e.d.r. directions for some estimate $1 \leq \widehat{\mathcal{K}} < p$ of $\mathcal{K}$. The estimated e.d.r. directions correspond to eigenvectors of the matrix $\widetilde{V}$ from (4.29) because Σ must be estimated by $\widehat{\Sigma}$.

We shall use the k-replacement scheme from (4.39) in Section 4.3.2 to obtain a sequence of contaminated samples $(X, Y)_m^{n,k}$, $m \in \mathbb{N}$, as follows. Let $k = \widehat{\mathcal{K}} \leq H$ and,

following (4.39), replace the first observation (x_{1_h}, y_{1_h}) in each of the first $k = \widehat{\mathcal{K}}$ slices with $(\widetilde{x}_{1_h}, y_{1_h})$ for

$$\widetilde{x}_{1_h,m} = t_m \widehat{\beta}_h + \widetilde{v}_h, \quad \widetilde{v}_h \in \mathbb{R}^p, \quad h = 1, \dots, \widehat{\mathcal{K}}; \tag{4.49}$$

where $\{t_m\}_{m=1}^{\infty}$ is a positive real sequence such that $t_m \to \infty$ as $m \to \infty$. That is, we contaminate in each of the $\widehat{\mathcal{K}}$ directions $\{\widehat{\beta}_h\}_{h=1}^{\widehat{\mathcal{K}}}$ estimated from the uncontaminated sample. We shall also assume that (4.45) holds after setting $\mathcal{K} = \widehat{\mathcal{K}}$ and $k = \widehat{\mathcal{K}}$ in Definition 4.2. (More generally, we may assume that the uncontaminated data $(X, Y)^n$ satisfy the breakdown pattern in Definition 4.2 after setting $\mathcal{K} = \widehat{\mathcal{K}}$ and $k = \widehat{\mathcal{K}}$ in that definition; the choice of contaminating the first observation in each of the first $k = \widehat{\mathcal{K}}$ slices is just a special case of contamination in which the conditions of Definition 4.2 simplify to the condition (4.45).)

The contaminated sample $(X, Y)_m^{n,k}$, with $k = \widehat{\mathcal{K}}$, yields an estimate $\widehat{\Sigma}_m$ of Σ as well as a $p \times p$ matrix $\widetilde{V}_m$ from (4.29). This matrix $\widetilde{V}_m$ has ordered eigenvalues $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ and corresponding eigenvectors $\widehat{\beta}_{m,i}$, $i = 1, \dots, p$, which are all p contaminated e.d.r. directions.

We find an estimate $\widehat{\mathcal{K}}_{m,k}$ based on $(X, Y)_m^{n,k}$ to produce a contaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}} = \mathrm{span}(\widehat{\beta}_{m,1}, \dots, \widehat{\beta}_{m,\widehat{\mathcal{K}}_{m,k}})$, where again $k = \widehat{\mathcal{K}}$. However, unlike in the case where Σ is known (and $\mathcal{K}$ unknown) from Section 4.4.1, we have *no apparent way* to control the size of the contaminated eigenvalues $0 \leq \widehat{\lambda}_{m,p} \leq \cdots \leq \widehat{\lambda}_{m,1}$ when Σ is *unknown*. In Section 4.3.2 we developed some strong results on the behavior of *eigenvectors* $\widetilde{V}_m$ (i.e., the contaminated e.d.r. directions) under contamination when Σ is unknown, but none of these results indicate the behavior of the *eigenvalues* of $\widetilde{V}_m$ under contamination. In fact, the simulation study of Chapter 5 will show that, under various types of contamination (similar to what we consider here), contaminated eigenvalues of $\widetilde{V}_m$ may differ largely or very slightly from the eigenvalues of the uncontaminated matrix $\widetilde{V}$ from (4.29) used to produce estimates $\widehat{\mathcal{K}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$. That is, it is very difficult to control the contaminated estimate $\widehat{\mathcal{K}}_{m,k}$ of the reduction subspace dimension.

The best result on the effect of contamination when Σ and $\mathcal{K}$ are unknown can be

summarized as follows. With the $k = \widehat{\mathcal{K}}-$contamination scheme described above (i.e., in the direction of $\{\widehat{\beta}_h\}_{h=1}^{\widehat{\mathcal{K}}}$), we can construct a sequence of contaminated data sets $(X,Y)_m^{n,k}$, $m \in \mathbb{N}$ with corresponding $\mathbb{R}^p$-subspace estimates $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ where it holds that $P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} \approx 0_{p\times p}$ (i.e., $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ and $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ are nearly orthogonal) and $\dim(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}) \leq p - \widehat{\mathcal{K}}$ for large m. That is, the contaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ must be essentially a subset of the orthogonal complement of the uncontaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}$ for large m, which has rank $p - \widehat{\mathcal{K}}$. Of course, this result is not as strong as saying that $\dim(\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}) = p - \widehat{\mathcal{K}}$ in addition to $P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} \approx 0_{p\times p}$ for large m, which was possible in Section 4.4.1 (known Σ). Consequently, the contaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ might not correspond to the entire orthogonal complement of the uncontaminated subspace estimate $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}^{\perp}$ for large m. However, contamination can still drastically alter the subspace estimate intended by the uncontaminated data $(X,Y)^n$ by actually causing a subspace estimate orthogonal to $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$.

We summarize the result in the following theorem.

Theorem 4.4 *For a given data set $(X,Y)^n$ with unknown values of $\mathrm{E}(X) = \mu$, $\mathrm{Cov}(X) = \Sigma$, suppose the SIR procedure yields an estimated reduction subspace $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}} \subset \mathbb{R}^p$ with an estimated dimension $1 \leq \widehat{\mathcal{K}} < p$, where $H \leq p$ slices are used in SIR. Provided that the data $(X,Y)^n$ satisfy the breakdown pattern Definition 4.2 (setting $\mathcal{K} = \widehat{\mathcal{K}}$, $k = \widehat{\mathcal{K}}$ in this Definition) then, there exists a sequence $(X,Y)_m^{n,k}$, $m \in \mathbb{N}$, of contaminated data sets and associated subspace estimates $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}} \subset \mathbb{R}^p$, found by replacing $k = \widehat{\mathcal{K}}$ observations in $(X,Y)^n$, for which the following hold.*

(i) For some $N \in \mathbb{N}$, the contaminated estimate $\widehat{\mathcal{K}}_{m,k}$ of the dimension of $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ satisfies

$$\widehat{\mathcal{K}}_{m,k} \leq p - \widehat{\mathcal{K}}, \qquad m \geq N.$$

(ii) $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}$ is orthogonal to $\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}$ as $m \to \infty$, namely

$$\lim_{m \to \infty} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}} P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} = 0_{p \times p}.$$

Remark. If the uncontaminated data $(X,Y)^n$ produces an estimate $\widehat{\mathcal{K}} = 0$ (no relationship between X and Y), results (i) and (ii) in the above theorem still hold.

Proof. Because the data $(X, Y)^n$ satisfy the breakdown pattern Definition 4.2 (setting $\mathcal{K} = \widehat{\mathcal{K}}$, $k = \widehat{\mathcal{K}}$), WloG we assume that we may contaminate the first observation in each of the first $k = \widehat{\mathcal{K}}$ slices as in (4.49) so that (4.45) holds with $\mathcal{K} = \widehat{\mathcal{K}}$, $k = \widehat{\mathcal{K}}$.

Now repeating the proof of CASE I of Corollary 4.2 with $\widehat{\mathcal{K}}$ substituted for $\mathcal{K}$ and setting $k = \widehat{\mathcal{K}}$ in that proof, we find that part (ii) above follows.

To show part (i), we note that

$$\mathrm{tr}\left((\mathrm{I}_p - P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}})P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}\right) \leq \mathrm{tr}(\mathrm{I}_p - P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}}) = p - \mathrm{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}}) = p - \widehat{\mathcal{K}},$$

so that by the result in part(i) it holds that

$$
\begin{aligned}
\limsup_{m \to \infty} \widehat{\mathcal{K}}_{m,k} &= \limsup_{m \to \infty} \mathrm{rank}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}) \\
&= \limsup_{m \to \infty} \mathrm{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}) \\
&= \limsup_{m \to \infty} \mathrm{tr}\left(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}}P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}} + (\mathrm{I}_p - P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}})P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}\right) \\
&= \limsup_{m \to \infty} \mathrm{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}}P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}) + \limsup_{m \to \infty} \mathrm{tr}\left((\mathrm{I}_p - P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}})P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}\right) \\
&\leq \limsup_{m \to \infty} \mathrm{tr}(P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}}}P_{\widehat{\mathcal{B}}_{\widehat{\mathcal{K}}_{m,k}}}) + p - \widehat{\mathcal{K}} \\
&= 0 + p - \widehat{\mathcal{K}} = p - \widehat{\mathcal{K}}. \tag{4.50}
\end{aligned}
$$

For each $m \in \mathbb{N}$, define $b_m = \sup\{\widehat{\mathcal{K}}_{m,k}, \widehat{\mathcal{K}}_{m+1,k}, \widehat{\mathcal{K}}_{m+2,k}, \ldots\}$ and note that $\limsup_{m \to \infty} \widehat{\mathcal{K}}_{m,k} = \inf_m b_m$ by definition; see the Appendix. Suppose we make the assumption that for all $m \in \mathbb{N}$, it holds that $b_m \geq p - \widehat{\mathcal{K}} + 0.5$ in which case $\limsup_{m \to \infty} \widehat{\mathcal{K}}_{m,k} = \inf_m b_m \geq p - \widehat{\mathcal{K}} + 0.5$; this is a contradiction of (4.50). Hence, it must be the case that there exists an $N \in \mathbb{N}$ such that $\sup\{\widehat{\mathcal{K}}_{N,k}, \widehat{\mathcal{K}}_{N+1,k}, \widehat{\mathcal{K}}_{N+2,k}, \ldots\} = b_N < p - \widehat{\mathcal{K}} + 0.5$ so that $\widehat{\mathcal{K}}_{m,k} < p - \widehat{\mathcal{K}} + 0.5$ for $m \geq N$. Because $\widehat{\mathcal{K}}_{m,k} \in \mathbb{N}$ is an integer, we have $\widehat{\mathcal{K}}_{m,k} \leq p - \widehat{\mathcal{K}}$ for $m \geq N$ which shows part (i) of the theorem. $\square$

Chapter 5

A Simulation Study

Our objective of Chapter 4 was a thorough investigation of the finite sample behavior of SIR under the presence of some type of data contamination. An important conclusion to be drawn from the results obtained during this investigation can be summarized as follows. Not every type of data contamination is necessarily damaging the estimation of the e.d.r. subspace. As it turns out, the knowledge of the covariance matrix Σ and of the dimension $\mathcal{K}$ of the e.d.r. subspace profoundly determine how harmful a certain type of data contamination essentially is for the e.d.r. subspace estimation. While one type of contamination causes estimates of e.d.r. directions orthogonal to the actual e.d.r. directions when Σ is known, this type of contamination will not effect the estimates when Σ is unknown and vice versa.

We want to proceed now with a simulation study to support our theoretical findings from Chapter 4. In particular, we wish to verify our results regarding the amount and the type of data contamination that is necessary to cause e.d.r. directions to be estimated orthogonal to the actual e.d.r. directions.

5.1 Simulation Design

We will look at two different statistical models. Both models have been investigated in similar form before: once in the dissertation of Hilker (1997, p.171), and also in the dissertation of Bond (1999, p.108). We assume the regressor space to be four-dimensional $X = (X_1, X_2, X_3, X_4)^\top$, where each regressor variable represents an independent observation from a standard normal distribution (i.i.d. $X_i \sim \mathcal{N}(0, 1), i = 1, \ldots, 4$). Note that the mean and covariance of the regressor variables X are given by $\mathrm{E}(X) = \mu = 0_4$ and $\mathrm{Cov}(X) = \Sigma = \mathrm{I}_4$. Observations for the response variable Y are determined with respect to Model 1 and Model 2 given below.

Model 1:
$$Y = X_1 + X_2 + X_3 + X_4. \tag{5.1}$$

Hence, we have $\mathcal{K} = 1$ with e.d.r. direction $\beta_1 = (1, 1, 1, 1)^\top$.

Model 2:
$$Y = \frac{X_1}{0.5 + \sqrt{|1.5 + X_2|}}. \tag{5.2}$$

Hence, we have $\mathcal{K} = 2$ with e.d.r. directions $\beta_1 = (1, 0, 0, 0)^\top$ and $\beta_2 = (0, 1, 0, 0)^\top$.

The corresponding e.d.r. subspaces for Model 1 and Model 2 are one- and two-dimensional, $\mathcal{K} = 1$ and $\mathcal{K} = 2$, respectively. For simplicity, we neglected an additional error term ε (see (3.1)) in both models. From each model, we consider taking samples $(X, Y)^n \in \mathbb{R}^4 \times \mathbb{R}$ of size $n = 100$, where some of these points will be replaced with corrupted values. We will consider applying SIR to contaminated samples using $H = 10$ slices, $\mathbb{I}_h, h = 1, \ldots, H$.

As described in Sections 4.3.1 and 4.3.2, contamination involving corrupted observations spread out in different slices can be damaging SIR. For both Models 1 and 2, we consider various amounts of contamination by replacing either one, two or three observations in a generated data set $(X, Y) \in \mathbb{R}^4 \times \mathbb{R}$ with contaminated x-values. To

obtain three replacement-contamination points, we shall replace the x-value of the first x-observation x_{1_h} in each of the first three slices according to the following structure

$$\widetilde{x}_{1_h,m} = t_m \widetilde{\beta}_h, \qquad \widetilde{\beta}_h \in \mathbb{R}^4, \quad h = 1, 2, 3,$$

where $\widetilde{\beta}_h$ denotes the direction of contamination and $t_m > 0$ is a selected scaling factor. To obtain one, two or three contamination points, we may use $\{\widetilde{x}_{1_1,m}\}$, $\{\widetilde{x}_{1_1,m}, \widetilde{x}_{1_2,m}\}$ or $\{\widetilde{x}_{1_1,m}, \widetilde{x}_{1_2,m}, \widetilde{x}_{1_3,m}\}$. We continue by precisely stating possible directions $\widetilde{\beta}_h$ of contamination for Models 1 and 2.

DIRECTION OF CONTAMINATION WHEN Σ IS UNKNOWN:

When Σ is unknown and has to be estimated, contaminated data points affect potential e.d.r. direction estimates if the contamination is placed in the direction of the e.d.r. directions itself; see Section 4.4.2. For Model 1 this corresponds to a contamination in the direction of

$$\widetilde{\beta}_1 = \beta_1 = (1, 1, 1, 1)^\top.$$

Hence, the replacement of **one** observation $x_{1_1,m}$ in one slice by $\widetilde{x}_{1_1,m}$ is sufficient in order to estimate β_1 orthogonal to itself.

The e.d.r. subspace for Model 2 is spanned by the vectors

$$\beta_1 = (1, 0, 0, 0)^\top \quad \text{and} \quad \beta_2 = (0, 1, 0, 0)^\top.$$

Consequently, effective contamination leading to estimates orthogonal to both β_1 and β_2 is possible if **two** observations $x_{1_1,m}, x_{1_2,m}$ in two distinct slices are replaced by $\widetilde{x}_{1_1,m}, \widetilde{x}_{1_2,m}$, respectively, using contamination $\widetilde{\beta}_h = \beta_h$, $h = 1, 2$, in the directions of the two actual e.d.r. directions.

DIRECTION OF CONTAMINATION WHEN Σ IS KNOWN:

For simplicity, we are assuming that the covariance matrix Σ corresponds to the identity matrix I_4. Effective contamination now has to involve orthonormal

contamination directions that are orthogonal to the actual e.d.r. directions in order to estimate directions orthogonal to the true e.d.r directions. Because the e.d.r. subspace of Model 1 is of dimension one, the orthogonal complement of $\mathrm{span}(\beta_1) \subset \mathbb{R}^4$ is of dimension three. By the results in Section 4.4.1, SIR will yield a three-dimensional e.d.r. subspace orthogonal to $\mathrm{span}(\beta_1)$ if we replace **three** distinct observations $x_{1_1,m}, x_{1_2,m}, x_{1_3,m}$ in three different slices with contaminated versions $\widetilde{x}_{1_1,m}, \widetilde{x}_{1_2,m}, \widetilde{x}_{1_3,m}$; the contaminated points should involve three different contamination directions $\{\widetilde{\beta}_h\}_{h=1}^3$ which are orthonormal and also orthogonal to β_1. Three orthonormal vectors orthogonal to β_1 are for example given by

$$\widetilde{\beta}_1 = \left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0 \right)^\top, \quad \widetilde{\beta}_2 = \left(0, 0, -\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right)^\top$$

and

$$\widetilde{\beta}_3 = \left(\frac{1}{\sqrt{4}}, \frac{1}{\sqrt{4}}, -\frac{1}{\sqrt{4}}, -\frac{1}{\sqrt{4}} \right)^\top.$$

Hence, contamination of three distinct observations, one for each of the above directions $\widetilde{\beta}_1, \widetilde{\beta}_2$ and $\widetilde{\beta}_3$ and each placed in a different slice is sufficient in order to estimate the complete orthogonal complement of $\mathrm{span}(\beta_1) = \mathrm{span}((1,1,1,1)^\top)$. By contaminating only one or two observations according to $\widetilde{\beta}_1$ or $\widetilde{\beta}_2$, only parts of the orthogonal complement will be estimated.

For Model 2, vectors that are orthonormal to the e.d.r. directions $\beta_1 = (1,0,0,0)^\top$ and $\beta_2 = (0,1,0,0)^\top$ are given by

$$\widetilde{\beta}_1 = (0,0,1,0)^\top \quad \text{and} \quad \widetilde{\beta}_2 = (0,0,0,1)^\top.$$

By contaminating **two** distinct observations $\{\widetilde{x}_{1_h,m}\}_{h=1}^2$ in two different slices using $\widetilde{\beta}_1$ and $\widetilde{\beta}_2$, respectively, SIR will estimate the orthogonal complement of the e.d.r. subspace spanned by β_1 and β_2.

Furthermore, the magnitude of contamination, denoted by t_m is also of interest to us. For fixed directions of contamination $\widetilde{\beta}_h$, we shall increase the magnitude t_m WLoG according to the following values

$$t_m = 10^0, \ t_m = 10^1, \ t_m = 10^2, \ t_m = 10^4, \ t_m = 10^6 \text{ and } t_m = 10^8.$$

For each model we will again vary the amount of contamination. For example, when $\Sigma = I_4$ is known, contamination for Model 1 in one, two or three directions is meaningful, while for Model 2 contamination in at most two directions has an important effect for SIR.

PERFORMANCE CRITERIA:

After replacing observations in a generated data set $(X, Y) \in \mathbb{R}^4 \times \mathbb{R}$ with one, two or three contaminated points $\{\widetilde{x}_{1_h,m}\}_{h=1}^3$, an application of SIR will yield estimated e.d.r directions $\widehat{\beta}_1, \ldots, \widehat{\beta}_4$ with corresponding eigenvalues $0 \leq \widehat{\lambda}_4 \leq \cdots \leq \widehat{\lambda}_1$. To assess the quality of these estimated e.d.r. directions under contamination, we evaluate the vector product $\widehat{\beta}_i^\top \beta_1$ for $i = 1, \ldots, 4$, (and in addition $\widehat{\beta}_i^\top \beta_2$ for Model 2). Because we normalized the estimated e.d.r. directions as well as the actual e.d.r. directions, these vector products correspond to the cosine of the angle between them, i.e we compute as previously defined in (2.1)

$$\cos \theta = \cos(\widehat{\beta}_i^\top \angle \beta_j) = \widehat{\beta}_i^\top \beta_j \quad \text{for} \quad i = 1, \ldots, 4; j = 1, 2,$$

where $\|\widehat{\beta}_i\| = 1 = \|\beta_j\|$ The cosine of the angle should be close to 0 when both vectors are orthogonal to each other and approximately 1 or -1 when they span the same direction.

The reason for choosing $\cos \theta$ as a performance criteria in this simulation study rather than the previously introduced Frobenius norm-based metric is that by applying $\mathbb{F}$ to the subspace estimates, we obtain information on how far apart the *subspace* estimates are. Information on the behavior of individual e.d.r. directions however is lost. We feel that using $\cos \theta$ at this point will give more insight to the effect of contamination on individual e.d.r. directions apart from any definition of breakdown on subspaces.

For each model (Model 1 or 2), contamination amount (one, two or three points from $\{\widetilde{x}_{1_h,m}\}_{h=1}^3$), and magnitude t_m, we conducted $M = 1000$ simulation runs in which we generated data sets $(X, Y)^n$ of sample size $n = 100$, contaminated the data and

computed $\widehat{\beta}_i^\top \beta_j = \cos(\widehat{\beta}_i^\top \angle \beta_1)$, $i = 1, \ldots, 4$ (and in addition $\cos(\widehat{\beta}_i^\top \angle \beta_2)$ for Model 2); the values of $|\cos(\widehat{\beta}_i^\top \angle \beta_1)|$ and $|\cos(\widehat{\beta}_i^\top \angle \beta_2)|$ were then averaged over the $M = 1000$ runs and reported in the subsequent tables. Using the absolute value makes sense as the cosine of the angle can take on the value 1 when vectors span the same direction. Average values of the ordered eigenvalues $\{\widehat{\lambda}_i\}_{i=1}^4$ were also tabulated.

For purposes of comparison, we also computed average values of $|\widehat{\beta}_i^\top \beta_1|$, $|\widehat{\beta}_i^\top \beta_2|$, and $\widehat{\lambda}_i$ ($i = 1, \ldots, 4$) for *uncontaminated* data sets $(X, Y)^n$ of sample size $n = 100$, using $M = 1000$ simulation runs.

5.2   Simulation results when Σ is unknown

We will proceed next with a summary of the most important results obtained in the simulation study when the covariance matrix Σ is unknown. The main results for both Models are numerically summarized in Tables 2 through 6.

- Table 2 displays the results of the simulation study for Model 1. We contaminated the first observation of the first slice in the direction of $\beta_1 = (1, 1, 1, 1)^\top$. Evidently, the contamination has an effect on all estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$, although this effect is strongest for $\widehat{\lambda}_1$; the only significant eigenvalue for Model 1. As we increase the magnitude of the contamination by letting t_m tend to larger values, we can observe that all estimated eigenvalues converge and remain approximately unchanged. This implies that the magnitude t_m has a relatively small effect on the eigenvalues compared to the direction of contamination β_1. This agrees with results in Section 4.4.2 that the size of contaminated estimated eigenvalues $\widehat{\lambda}_i$ are difficult to control via the magnitude of contamination t_m.

- With respect to the estimated e.d.r. directions we find in Table 2, that SIR correctly estimates the true e.d.r. direction $\beta_1 = (1, 1, 1, 1)^\top$ with $\widehat{\beta}_1$ when there is no contamination. However, by the contamination of one observation, SIR is not

able to correctly estimate β_1 any longer. Not only is the estimate of the first e.d.r. direction, $\widehat{\beta}_1$, almost orthogonal to β_1, i.e. $|\cos(\widehat{\beta}_1^\top \angle \beta_1)| \approx 0.0204$ for $t_m = 10^2$, the remaining directions $\widehat{\beta}_2, \widehat{\beta}_3$ and $\widehat{\beta}_4$ are estimated orthogonally to β_1 as well. This essentially supports the finding in Theorem 4.4(ii); an estimated e.d.r. space under contamination will be orthogonal to the direction of contamination β_1 (or the space spanned by the uncontaminated estimate $\widehat{\beta}_1$). Furthermore we can see that, contrary to the eigenvalues, the magnitude of contamination t_m yields estimates for $\beta_1, \ldots, \beta_4$ that tend to be increasingly orthogonal to β_1 and thus the magnitude has an effect on the e.d.r. directions.

Therefore, a legitimate conclusion is, that due to the contamination of one observation in the direction of $\beta_1 = (1, 1, 1, 1)^\top$, Sir is not able to recover the true e.d.r. direction β_1 for Model 1 any longer.

- Tables 3 to 6 summarize our findings obtained for Model 2. In Table 3 the results for the estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$ are displayed when we contaminated exactly one observation in the direction of $\beta_1 = (1, 0, 0, 0)^\top$ and a second in the direction of $\beta_2 = (0, 1, 0, 0)$, each placed in a different slice. The results are similar to the ones obtained before for Model 1. Again we find that the largest eigenvalue $\widehat{\lambda}_1$ shows the greatest amount of change under contamination and appears to decrease the most relative to the uncontaminated value of $\widehat{\lambda}_1$. Obviously the contamination itself does have an effect on the estimated eigenvalues, the magnitude t_m however does not, as the average eigenvalue estimates remain fairly unchanged for increasing values of t_m. Once more, the size of the estimated eigenvalues under contamination are difficult to directly control through the magnitude of contamination.

- In Table 4 we contaminated only one observation. The left side of the table shows the estimated eigenvalues when we contaminated the first observation of the first slice in the direction of β_1, while on the right hand side, results are presented when the contamination of the first observation of the first slice was placed in the direction of β_2. An interesting observation to be made here is the

following. We find in Table 4 that a contamination in the direction of β_1 results in a contamination of all estimated eigenvalues $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_4$ as they all decrease in value and similar to what we have seen earlier (where the corrupted estimate $\widehat{\lambda}_1$ again exhibits the most change). However, a contamination in the direction of β_2 causes very little change in $\widehat{\lambda}_1$ while the remaining estimates $\widehat{\lambda}_2, \ldots, \widehat{\lambda}_4$ decrease only slightly. This may seem somewhat surprising in the sense that the behavior of estimated eigenvalues from SIR under contamination is not *symmetric* with respect to the direction of contamination. Hence, the direction of contamination in Model 2 seems to play a role in determining eigenvalue estimates.

- Tables 5 and 6 outline the results on the estimated e.d.r. directions under contamination corresponding to the eigenvalues found in Tables 3 and 4. Without contamination, SIR appears to correctly estimate β_1 but is somewhat less successful in estimating second true e.d.r. direction β_2. The contamination of two observations in the directions of β_1 and β_2, respectively caused SIR to estimate all directions orthogonal to β_1 and β_2. That is, none of the estimated e.d.r. directions, regardless of the significance of the corresponding eigenvalues, are elements of the true e.d.r. subspace spanned by β_1 and β_2; this again supports the theoretical findings in Section 4.4.2.

  In Table 6 we displayed numerical results on the e.d.r direction estimates obtained from using one contaminated observation 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 4), the procedure could also recover β_1 as well through the contaminated estimate $\widehat{\beta}_1$ (i.e., the absolute value of the cosine of the angle $|\cos(\widehat{\beta}_1^\top \angle \beta_1)|$ assumed values always fairly close to one). More interesting to us are the results on the right hand side of Table 6. Because of the contamination in the direction 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.

The conclusion we can draw from this first part of our simulation study where we assumed Σ to be unknown, is that the findings support the theory established in

Section 4. In particular we were able to verify the claims made in Theorem 4.2 and Theorem 4.4. When the covariance matrix Σ is unknown, we are not really able to control the size of the contaminated eigenvalues from the contaminated covariance matrix $\widetilde{V}_m$ version of (4.29) through the magnitude of contamination t_m. While explicit control of estimated e.d.r. directions can be made through contamination (i.e. the eigenvectors of $\widetilde{V}_m$), we cannot make general statements regarding the contaminated eigenvalues from SIR under the same form of contamination as contaminated eigenvalues may differ only slightly from their uncontaminated counterparts.

CONTAMINATION	ESTIMATED EIGENVALUES				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$	$\lvert\cos(\widehat{\beta}_1^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_2^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_3^\top \angle \beta_1)\rvert$	$\lvert\cos(\widehat{\beta}_4^\top \angle \beta_1)\rvert$
$t_m = 10^0$	0.84270409	0.14000137	0.07288239	0.03052472	0.9988623	0.0882621	0.08934124	0.09066589
$t_m = 10^1$	0.18622889	0.11534647	0.06547804	0.02821972	0.357556	0.2856594	0.1780135	0.1429609
$t_m = 10^2$	0.16833012	0.10077923	0.05791316	0.02455521	0.02041144	0.03534819	0.03754557	0.02660068
$t_m = 10^4$	0.17343339	0.10487489	0.05887939	0.02532071	0.0002624112	0.0003902502	0.0003089088	0.0002232264
$t_m = 10^6$	0.17048452	0.10328906	0.05955203	0.02631496	2.779502e-06	3.926096e-06	3.236469e-06	2.179378e-06
$t_m = 10^8$	0.17571454	0.10533471	0.06018917	0.02637326	2.556765e-08	3.958522e-08	3.193362e-08	2.192166e-08
NO CONTAMINATION	0.96199782	0.14268208	0.07262518	0.03028078	0.9992103	0.07768901	0.08401958	0.08068395

TABLE 2.   Model 1 – Behavior of eigenvalues and e.d.r. directions under contamination of one data point in the direction of $\beta_1 = (1, 1, 1, 1)^\top$, Σ unknown

TABLE 3. Model 2 – Behavior of eigenvalues under contamination of two data points – one in the direction of $\beta_1 = (1,0,0,0)^\top$ and one in the direction of $\beta_2 = (0,1,0,0)^\top$, Σ unknown

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
$t_m = 10^8$	0.16299350	0.10462858	0.06583942	0.02941150
NO CONTAMINATION	0.92968171	0.18887460	0.08852808	0.03577398

TABLE 4. Model 2 – Behavior of eigenvalues under contamination of only one data point in the direction of $\beta_1 = (1,0,0,0)^\top$ (LEFT) or in the direction of $\beta_2 = (0,1,0,0)^\top$ (RIGHT), Σ unknown

CONTAMINATION	ESTIMATED EIGENVALUES			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	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
$t_m = 10^8$	0.21130982	0.11599203	0.06273975	0.02757866
NO CONTAMINATION	0.92949550	0.18676763	0.08929240	0.03601315

CONTAMINATION	ESTIMATED EIGENVALUES			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	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
$t_m = 10^8$	0.93160782	0.13159777	0.06801507	0.02965038
NO CONTAMINATION	0.92919865	0.18293314	0.08698891	0.03665909

CONTAMINATION	COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$|\cos(\widehat{\beta}_1^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_2^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_3^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_4^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_1^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_2^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_3^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_4^\top \angle\beta_2)|$
$t_m = 10^0$	0.9973476	0.0972353	0.09455044	0.09119261	0.04509871	0.7150232	0.4017536	0.3328404
$t_m = 10^1$	0.8950354	0.1805444	0.1802694	0.1612251	0.3173347	0.3630845	0.4169581	0.3783862
$t_m = 10^2$	0.05231419	0.08107545	0.1117936	0.0875233	0.07831053	0.1022642	0.08702715	0.06199629
$t_m = 10^4$	0.0006141142	0.001902911	0.001151127	0.0004754775	0.0006027861	0.001912108	0.001109116	0.0004811929
$t_m = 10^6$	5.428481e-06	1.862893e-05	1.235388e-05	4.690284e-06	5.428294e-06	1.890622e-05	1.218958e-05	4.619279e-06
$t_m = 10^8$	5.531661e-08	1.859097e-07	1.166303e-07	4.53728e-08	5.635411e-08	1.851129e-07	1.143412e-07	4.60066e-08
NO CONTAMINATION	0.9981598	0.08652421	0.08502864	0.0850214	0.07831053	0.1022642	0.08702715	0.06199629

TABLE 5. Model 2 – Behavior of e.d.r. directions under contamination of two data points – one in the direction of $\beta_1 = (1,0,0,0)^\top$ and one in the direction of $\beta_2 = (0,1,0,0)^\top$, Σ unknown

CONTAMINATION	COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$|\cos(\widehat{\beta}_1^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_2^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_3^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_4^\top \angle\beta_1)|$	$|\cos(\widehat{\beta}_1^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_2^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_3^\top \angle\beta_2)|$	$|\cos(\widehat{\beta}_4^\top \angle\beta_2)|$
$t_m = 10^0$	0.997704	0.09540198	0.08763669	0.0851936	0.0409813	0.7044577	0.416999	0.3324463
$t_m = 10^1$	0.9947148	0.1068113	0.1190944	0.1354513	0.0903446	0.2760754	0.3938079	0.5353305
$t_m = 10^2$	0.9988849	0.1162464	0.1884353	0.1964554	0.01692782	0.04406396	0.08927996	0.09609494
$t_m = 10^4$	0.9989754	0.128747	0.2022079	0.1719973	0.0001718877	0.000515331	0.001011452	0.0008247947
$t_m = 10^6$	0.9989905	0.1268841	0.2079789	0.1703580	1.723068e-06	5.011812e-06	1.008949e-05	8.171434e-06
$t_m = 10^8$	0.9990158	0.1257864	0.2061445	0.1752891	1.729672e-08	4.947694e-08	1.019937e-07	8.289734e-08
NO CONTAMINATION	0.9980909	0.08136593	0.08509767	0.08603543	0.03353829	0.7515609	0.3812316	0.3143829

TABLE 6. Model 2 – Behavior of e.d.r. directions under contamination of only one data point in the direction of $\beta_2 = (0,1,0,0)^\top$, Σ unknown

5.3   Simulation results when Σ is known

Under the assumption that the covariance matrix $\Sigma = I_4$ is known and this is accounted for when applying the SIR procedure, the conducted simulation study yielded the following results numerically summarized in Tables 7 through 12.

- Tables 7 through 9 display the results that we obtained for Model 1. We started with investigating the behavior of SIR under the contamination of exactly one observation, where we chose again WLoG the first observation of the first slice and contaminated it in the direction of $\widetilde{\beta}_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)^\top$, which is orthonormal with respect to the true e.d.r. direction $\beta_1 = (1, 1, 1, 1)^\top$ (see Table 7). A main result from Chapter 4 was that for the case when Σ is known, a contamination of k slices causes the k largest eigenvalues of the contaminated matrix $\widehat{V}_m$ version of (4.20) to "explode" and to grow infinitely large at a rate faster than the remaining $p - k$ eigenvalues of $\widehat{V}_m$; see Theorem 4.1(b) and the subsequent remark to Theorem 4.1. The under contamination estimated eigenvalues in Table 7 clearly seem to support this finding. 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 three eigenvalues $\widehat{\lambda}_2$, $\widehat{\lambda}_3$ and $\widehat{\lambda}_4$. The same effect can be observed in Table 8 and Table 9, where we contaminated two and three observations, respectively. In Table 8, we contaminated the first observation of the first slice in the direction of $\widetilde{\beta}_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)^\top$ and the first observation of the second slice in the direction of $\widetilde{\beta}_2 = (0, 0, -1/\sqrt{2}, -1/\sqrt{2})^\top$. Now, the estimated first two eigenvalues $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ grow infinitely large as t_m increases while $\widehat{\lambda}_3$ and $\widehat{\lambda}_4$ remain bounded. Table 9 differs from the previous two tables only with respect to an additionally contaminated observation replaced in the third slice in the direction of $\widetilde{\beta}_3 = (1/\sqrt{4}, 1/\sqrt{4}, -1/\sqrt{4}, -1/\sqrt{4})^\top$. The effect on the eigenvalues is essentially the same. In addition to $\widehat{\lambda}_1$ and $\widehat{\lambda}_2$ we have now also $\widehat{\lambda}_3$ growing infinitely large as t_m increases.

- The right hand sides of Tables 7 through 9 display the results regarding the estimated e.d.r. directions under contamination. In each table, as more contaminated points are used (involving orthonormal directions of contamination with respect to β_1), we find that contaminated direction estimates $\widehat{\beta}_i$ corresponding to exploding eigenvalues $\widehat{\lambda}_i$ are orthogonal to the true e.d.r. direction β_1. This supports our findings in Theorem 4.1(c), which stated that eigenvectors associated with the k largest eigenvalues of the contaminated covariance matrix $\widehat{V}_m$ (under k-replacement contamination) converge to vectors spanned by the k (orthonormal) directions of contamination (i.e., orthogonal to the true e.d.r. subspace $\text{span}(\beta_1)$).

- The results for Model 2 are summarized in Tables 10 through 12. In Table 10, we display the contaminated eigenvalue estimates after contaminating two observations (WLoG, the first observation in the first two slices) in the directions $\widetilde{\beta}_1 = (0,0,1,0)^\top$, $\widetilde{\beta}_2 = (0,0,0,1)^\top$ which are orthonormal with respect to the two true e.d.r directions $\beta_1 = (1,0,0,0)^\top$, $\beta_2 = (0,1,0,0)^\top$ in Model 2. As the magnitude of contamination t_m increase, we find that the two largest contaminated eigenvalues explode in value because of the contamination of two data points; the smallest two eigenvalues remain bounded in value. Table 10 also shows that, if only one contamination point is used in the direction $\widetilde{\beta}_1 = (0,0,1,0)^\top$, then only one (the largest) contaminated eigenvalue explodes in size, while the remaining three contaminated eigenvalues are much smaller and essentially bounded.

Table 11 shows that, when two directions $\widetilde{\beta}_1 = (0,0,1,0)^\top$, $\widetilde{\beta}_2 = (0,0,0,1)^\top$ of contamination are used, the contaminated direction estimates $\widehat{\beta}_1, \widehat{\beta}_2$ corresponding to the two largest (exploding) contaminated eigenvalues are determined by $\widetilde{\beta}_1, \widetilde{\beta}_2$ and so are orthogonal to the two true e.d.r directions $\beta_1 = (1,0,0,0)^\top$, $\beta_2 = (0,1,0,0)^\top$ in Model 2. This would imply that a contaminated subspace estimate $\text{span}(\widehat{\beta}_1, \widehat{\beta}_2)$, if determined by the two clearly significant contaminated eigenvalues, would be orthogonal to the true e.d.r. subspace $\text{span}(\beta_1, \beta_2) \subset \mathbb{R}^4$.

Table 12 shows that, if one direction $\widetilde{\beta}_1 = (0,0,1,0)^\top$ of contamination is

used, then this determines the direction of contaminated direction estimate $\widehat{\beta}_1$ corresponding to the largest, exploding contaminated eigenvalue given in Table 10. Consequently, the two true e.d.r. directions β_1, β_2 are orthogonal to $\text{span}(\widehat{\beta}_1) \approx \text{span}((0, 0, 1, 0)^\top)$.

CONTAMINATION	ESTIMATED EIGENVALUES				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$	$|\cos(\widehat{\beta}_1^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_2^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_3^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_4^\top \angle \beta_1)|$
$t_m = 10^0$	0.89943299	0.14011246	0.06967781	0.02807165	0.979841	0.09456183	0.09123813	0.09043206
$t_m = 10^1$	0.93754552	0.17523492	0.08204017	0.03292336	0.9591784	0.1581040	0.1250104	0.1117938
$t_m = 10^2$	9.35701843	0.62411524	0.10242710	0.03875058	0.1750000	0.964122	0.1140435	0.1086621
$t_m = 10^4$	8.999768e+04	6.375259e-01	1.011727e-01	3.766357e-02	0.001672263	0.9797691	0.1204768	0.1043951
$t_m = 10^6$	9.000002e+08	6.362138e-01	1.006758e-01	3.855718e-02	1.666798e-05	0.9788298	0.1198380	0.1119880
$t_m = 10^8$	9.000000e+12	6.430530e-01	1.007416e-01	3.762832e-02	1.673450e-07	0.9804973	0.1156420	0.1079803
NO CONTAMINATION	0.98671578	0.14187920	0.07006886	0.02914564	0.981705	0.09123825	0.0871689	0.08483905

TABLE 7.   Model 1 – Behavior of eigenvalues and e.d.r. directions under contamination of only one data point in the direction of
$\widetilde{\beta}_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)^\top$, Σ known

CONTAMINATION	ESTIMATED EIGENVALUES				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$	$|\cos(\widehat{\beta}_1^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_2^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_3^\top \angle \beta_1)|$	$|\cos(\widehat{\beta}_4^\top \angle \beta_1)|$
$t_m = 10^0$	0.87849150	0.13586764	0.06707110	0.02734312	0.9796738	0.0946602	0.09170605	0.08955545
$t_m = 10^1$	0.92174541	0.21847900	0.10502371	0.04010948	0.9468494	0.1469614	0.1685776	0.1377121
$t_m = 10^2$	10.23292101	8.40767119	0.43498218	0.05851056	0.06132069	0.2108020	0.9572554	0.1432630
$t_m = 10^4$	9.999739e+04	8.000031e+04	4.515726e-01	5.869024e-02	0.0003923886	0.002162707	0.985721	0.1317057
$t_m = 10^6$	1.000000e+09	7.999998e+08	4.545222e-01	6.051161e-02	3.972363e-06	2.177360e-05	0.9849557	0.138369
$t_m = 10^8$	1.000000e+13	8.000000e+12	4.523386e-01	5.846076e-02	3.967582e-08	2.17854e-07	0.9851783	0.1364203
NO CONTAMINATION	0.97802389	0.14168882	0.06963360	0.02806297	0.980554	0.09614678	0.0887045	0.08595423

TABLE 8.   Model 1 – Behavior of eigenvalues and e.d.r. directions under contamination of only two data points in the directions of
$\widetilde{\beta}_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)^\top$ and $\widetilde{\beta}_2 = (0, 0, -1/\sqrt{2}, 1/\sqrt{2})^\top$, Σ known

CONTAMINATION	ESTIMATED EIGENVALUES				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$	$|\cos(\widehat{\beta}_1^{\top} \angle \beta_1)|$	$|\cos(\widehat{\beta}_2^{\top} \angle \beta_1)|$	$|\cos(\widehat{\beta}_3^{\top} \angle \beta_1)|$	$|\cos(\widehat{\beta}_4^{\top} \angle \beta_1)|$
$t_m = 10^0$	0.85937553	0.13956426	0.06805860	0.02836590	0.978474	0.1010335	0.09293968	0.09020317
$t_m = 10^1$	0.9229757	0.2560026	0.1349992	0.0569284	0.9451667	0.1352988	0.1440981	0.1852461
$t_m = 10^2$	10.706793	9.648394	7.447724	3.02329e-01	0.05339378	0.0584024	0.2467157	0.9640377
$t_m = 10^4$	1.000544e+05	9.994727e+04	6.996994e+04	3.216579e-01	0.0004063888	0.0004063516	0.002578214	0.9999964
$t_m = 10^6$	1.000005e+09	9.999947e+08	7.000002e+08	3.254403e-01	4.14553e-06	3.837078e-06	2.576043e-05	1
$t_m = 10^8$	1.000000e+13	9.999999e+12	7.000000e+12	3.227648e-01	4.0585e-08	4.068115e-08	2.583330e-07	1
NO CONTAMINATION	0.98090816	0.14086985	0.07032647	0.02850672	0.980692	0.09342744	0.09005521	0.08810476

TABLE 9. Model 1 – Behavior of eigenvalues and e.d.r. directions under contamination of three data points in the directions of $\widetilde{\beta}_1 = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0)^{\top}$, $\widetilde{\beta}_2 = (0, 0, -1/\sqrt{2}, 1/\sqrt{2})^{\top}$ and $\widetilde{\beta}_3 = (1/\sqrt{4}, 1/\sqrt{4}, -1/\sqrt{4}, -1/\sqrt{4})^{\top}$, Σ known

CONTAMINATION	ESTIMATED EIGENVALUES				ESTIMATED EIGENVALUES			
	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$	$\widehat{\lambda}_1$	$\widehat{\lambda}_2$	$\widehat{\lambda}_3$	$\widehat{\lambda}_4$
$t_m = 10^0$	0.85883928	0.17248466	0.08269297	0.03242646	0.88645859	0.17307941	0.08045838	0.03250486
$t_m = 10^1$	0.90970086	0.24396222	0.12369783	0.04602674	0.90998691	0.21522300	0.09430058	0.03524888
$t_m = 10^2$	10.25584498	8.39918308	0.45666037	0.06829026	9.35296603	0.64353573	0.11250216	0.04144181
$t_m = 10^4$	1.000010e+05	8.000099e+04	4.721786e-01	6.882700e-02	9.000340e+04	6.468671e-01	1.164591e-01	4.208303e-02
$t_m = 10^6$	1.000000e+09	8.000000e+08	4.744523e-01	6.934296e-02	9.000003e+08	6.511996e-01	1.176982e-01	4.134651e-02
$t_m = 10^8$	1.000000e+13	8.000000e+12	4.689955e-01	6.926768e-02	9.000000e+12	6.510415e-01	1.170527e-01	4.352920e-02
NO CONTAMINATION	0.94362646	0.18633471	0.08340895	0.03252900	0.95722132	0.18075728	0.08211033	0.03155600

TABLE 10. Model 2 – Behavior of eigenvalues under contamination of two data points in the directions of $\widetilde{\beta}_1 = (0, 0, 1, 0)^{\top}$ and $\widetilde{\beta}_2 = (0, 0, 0, 1)^{\top}$ (LEFT) and only one data point in the direction of $\widetilde{\beta}_1 = (0, 0, 1, 0)^{\top}$ (RIGHT), Σ known

CONTAMINATION	COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\|\cos(\widehat{\beta}_1^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_2^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_3^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_4^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_1^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_2^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_3^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_4^\top \angle\beta_2)\|$
$t_m = 10^0$	0.976686	0.1044171	0.1003012	0.0930625	0.1076441	0.696756	0.4183349	0.3421495
$t_m = 10^1$	0.9462966	0.1389971	0.1479168	0.172176	0.1005299	0.4668585	0.5072171	0.5463847
$t_m = 10^2$	0.05638568	0.2089328	0.9243171	0.2715613	0.03351661	0.03859145	0.2849067	0.9440719
$t_m = 10^4$	0.0003398631	0.002153915	0.9544445	0.2576584	0.0003178814	0.0004210452	0.2576596	0.9544464
$t_m = 10^6$	3.436004e-06	2.159297e-05	0.9522677	0.2661071	3.228445e-06	4.161506e-06	0.2661071	0.9522677
$t_m = 10^8$	3.431121e-08	2.167513e-07	0.9535582	0.2635546	3.260335e-08	4.121597e-08	0.2635546	0.9535582
NO CONTAMINATION	0.9793994	0.09976394	0.09180281	0.08587604	0.09870707	0.7433515	0.3914814	0.3097636

TABLE 11.   Model 2 – Behavior of e.d.r. directions under contamination of two data points in the directions of $\widetilde{\beta}_1 = (0, 0, 0, 1)^\top$ and $\widetilde{\beta}_2 = (0, 0, 1, 0)^\top$, Σ known

CONTAMINATION	COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS				COSINE OF ANGLE BETWEEN E.D.R. DIRECTIONS			
	$\|\cos(\widehat{\beta}_1^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_2^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_3^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_4^\top \angle\beta_1)\|$	$\|\cos(\widehat{\beta}_1^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_2^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_3^\top \angle\beta_2)\|$	$\|\cos(\widehat{\beta}_4^\top \angle\beta_2)\|$
$t_m = 10^0$	0.978405	0.1023728	0.09459632	0.08974925	0.1018218	0.6900753	0.426078	0.3442041
$t_m = 10^1$	0.959678	0.1413878	0.1289569	0.1221896	0.1019382	0.5873918	0.501991	0.4263278
$t_m = 10^2$	0.1689639	0.9506923	0.1616271	0.1350173	0.04366342	0.1932651	0.7186343	0.5150956
$t_m = 10^4$	0.001607249	0.9670707	0.1565432	0.1363324	0.0004653387	0.1748069	0.7117204	0.5323181
$t_m = 10^6$	1.607607e-05	0.9671019	0.1584999	0.1303249	4.51442e-06	0.1893439	0.7258426	0.5147027
$t_m = 10^8$	1.609887e-07	0.9669588	0.1594572	0.1318611	4.513712e-08	0.1827859	0.7267347	0.5161708
NO CONTAMINATION	0.979916	0.09882539	0.08718784	0.08826487	0.09881312	0.733789	0.4000374	0.3181767

TABLE 12.   Model 2 – Behavior of e.d.r. directions under contamination of only one data point in the direction of $\widetilde{\beta}_1 = (0, 0, 1, 0)^\top$, Σ known

Chapter 6

Quantitative Concepts for Dimension Reduction Procedures of the SIR-Type

After studying the effects of contaminated data on SIR in the finite sample case in Chapter 4 and defining a version of a finite sample breakdown point for the SIR dimension reduction procedure, we will focus now on some more general issues regarding the robustness of SIR at the level of population distributions.

In assessing robustness of a statistical functional (an estimator or test statistic) such as SIR at the population distribution level, one distinguishes between qualitative and quantitative robustness. While qualitative robustness provides information about the robustness of a functional against some type of contamination in general, quantitative robustness measures the degree of robustness of the functional. We will discuss both types of robustness in the context of the SIR procedure but focus mainly on quantitative robustness.

Qualitative robustness of a functional is linked to the equicontinuity of the functional. This property can be regarded as a minimal robustness condition. It is somewhat limited, because it does not permit comparisons between different qualitatively

robust estimators. That is, qualitative robustness gives no indication of the level of robustness. We note that the expectation functional, for example, is not qualitatively robust [see Hampel et al. (1986)] so that many classical statistical procedures, based on the mean, must also lack the property of qualitative robustness. Because the SIR procedure is mean-based, we can expect SIR to fail to be qualitatively robust as well. In addition, Davies & Gather (2004, 2005b) discuss that Principal Components and hence the SIR functional which is based on covariance matrix estimates and principal components is not even a continuous functional. Consequently, one could suspect SIR to have limited robustness.

The breakdown point and the influence function of an estimator belong to the class of quantitative measures of robustness. The breakdown point was originally introduced by Hampel (1968, 1971) and later also by Huber (1981) and Donoho & Huber (1983) in a finite sample version. Also of importance is the influence function, established first by Hampel (1971). Many other common measures can be introduced, such as the maxbias curve, the gross-error sensitivity, and the asymptotic variance for example [cf. Hampel, et al. (1986)]. In addition, Hampel et al. (1986, p.99) point out that if a functional possesses a positive breakdown point, it will generally be also qualitatively robust.

The main points of Chapter 6 can be summarized as follows.

- In Section 6.1, we will focus on the breakdown point of SIR at the level of probability distribution contamination. However, this discussion also has relevance for the finite sample breakdown of SIR (i.e., contamination of the empirical distribution, say P_n, of a particular data set $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^p \times \mathbb{R}$). We shall review the recent work of Davies and Gather (2002, 2005a) which is crucial for understanding if a meaningful breakdown point can be formulated for a statistical functional, like SIR. In essence, Davies and Gather propose a framework for a statistical functional that is required in finding a non-trivial upper bound for the breakdown point. This framework includes, among other issues involving metrics, the existence of a *rich* group equivariance structure.

The most important finding of the results here is that a group equivariance structure does in fact exist for SIR, as shown in Section 6.1.1. However, the SIR functional and subspace estimation in general *does not fit* into the framework of Davies and Gather (2002, 2005a) because an unbounded metric seemingly cannot be defined on the parameter space in SIR (which consists of *vector subspaces* of $\mathbb{R}^p$). Consequently, a meaningful breakdown concept for SIR, involving a breakdown point with an upper bound of 1/2, *cannot* be obtained. The implication of this may be that the entire concept of breakdown has limited value for SIR.

- Section 6.2 discusses a new and alternative breakdown point concept of Davies and Gather (2004, 2005b) which can be successfully applied to SIR as well as *any* dimension reduction functional.

- For completeness, in Section 6.3, we review the recent work of Prendergast (2004) in developing the influence function of SIR. The findings of Prendergast (2004) are relevant here because these further confirm the finite sample breakdown behavior of SIR, presented in Chapter 4 as well as the work of Hilker (1997) and Gather, Hilker and Becker (2002).

6.1 Breakdown Set-up of Davies & Gather

The relationship between a concept of BREAKDOWN and the existence of a group structure was introduced by Davies and Gather (2002, 2005a). They show that under a certain framework involving groups and metrics, which we shall explore in more detail in the following, a nontrivial upper bound for the breakdown point of a functional $\mathcal{T}$ can be derived. Up to this point, their work constitutes by far the most insightful one on breakdown. However, as Davies and Gather (2002) note, there exist earlier references in the literature observing a connection between an underlying group equivariance structure of a functional to its robustness properties; see He and Simpson (1993).

To provide a general description of the breakdown results of Davies and Gather, we first provide the key ingredients beginning with the group structure. Let $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ be a measurable sample space and $\mathcal{P}$ be a family of non-degenerate probability measures on $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ that are of primary interest. Let Θ be some parameter space and $\mathcal{T} : (\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_\Theta) \longrightarrow \Theta$ a functional. Let G denote a group of measurable transformations on the sample space $\mathcal{X}$ with elements $g : \mathcal{X} \to \mathcal{X}$, where for any $P \in \mathcal{P}$ and any $g \in G$ we set $P^g = P^g(B) = P(g^{-1}(B))$. For $n \in \mathbb{N}, n \geq 2$ and $g \in G$, we define $g^n(\cdot) = g(g^{n-1}(\cdot))$. The group G of data transformations induces a group H_G acting on the parameter space Θ by $H_G = \{h_g : g \in G\}$ where $h_g : \Theta \to \Theta$ such that $\mathcal{T}$ is called equivariant w.r.t. G if and only if $\mathcal{T}(P^g) = h_g(\mathcal{T}(P))$.

The breakdown results of Davies and Gather also require two (pseudo)metrics, d and D, respectively defined on the spaces of probability distributions $\mathcal{P}$ and parameters Θ. We denote a pseudometric on $\mathcal{P}$ by d, i.e. $d : \mathcal{P} \times \mathcal{P} \longrightarrow [0, \infty)$, which should satisfy the following two properties

$$\sup_{P, Q \in \mathcal{P}_\Theta} d(P, Q) = 1 \tag{6.1}$$

and

$$d(\alpha P + (1-\alpha)Q_1, \alpha P + (1-\alpha)Q_2) \leq 1 - \alpha \tag{6.2}$$

for any $P, Q_1, Q_2 \in \mathcal{P}_\Theta$ and $\alpha \in (0, 1)$. Condition (6.2) is a technical condition (needed in the proof of Theorem 6.1 below). A suitable pseudometric on the parameter space Θ is given by D, where $D : \Theta \times \Theta \longrightarrow [0, \infty)$ should satisfy

$$\sup_{\theta_1, \theta_2 \in \Theta} D(\theta_1, \theta_2) = \infty. \tag{6.3}$$

Figure 1 below displays the connection between the equivariance structure of a functional $\mathcal{T}$ on the sample space and the induced equivariance structure on the parameter space, as introduced by Davies and Gather (2002).

Before proceeding further, we define a subset of G of special interest by

$$G_1 = \left\{ g \in G : \lim_{n \to \infty} \inf_\theta D\left(\theta, h_{g^n}(\theta)\right) = \infty \right\}. \tag{6.4}$$

Figure 6.1: Group and Parameter Equivariance Structures
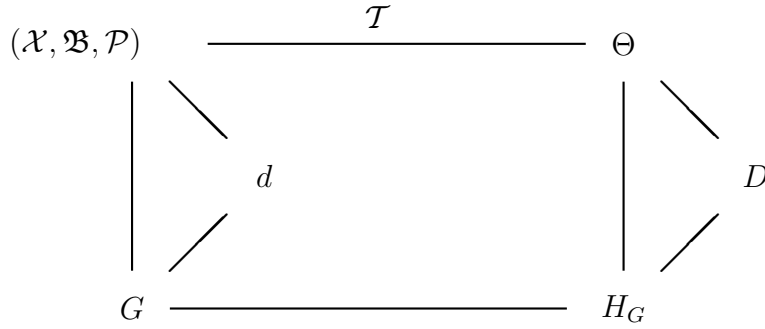
The collection G_1 is the set of all transformations $g \in G$ for which the induced transformation h_g causes θ and $h_{g^n}(\theta)$ to become maximally distant in the limit (i.e., upon iteratively applying a transformation g of the sample space $\mathcal{X}$ and the corresponding transformation h_g of the parameter space) with respect to the pseudometric D. We shall denote the restriction of a transformation $g \in G$ to a set $B \in \mathfrak{B}$ by $g_{|B}$ and define

$$\Delta(P) = \sup\{P(B) : B \in \mathfrak{B}, \ g_{|B} = \iota_{|B} \text{ for some } g \in G_1\}, \tag{6.5}$$

where $\iota \in G$ denotes an identity mapping, i.e. $\iota(x) = x$ for $x \in \mathcal{X}$. Hence, the quantity $\Delta(P)$ represents the largest probability measure of a set $B \in \mathfrak{B}$ for which some $g \in G_1$ from (6.4) maps the elements of B to itself, like the identity function.

For clarity in our notation and discussion, we repeat the breakdown point definitions of Davies and Gather (2002, 2005a).

Definition 6.1 *(**Davies & Gather, 2002, 2005a**) The **breakdown point** $\varepsilon^*(\mathcal{T}, P, d, D)$ of a functional $\mathcal{T}$ at a distribution P with respect to a pseudometrics d and D is defined by*

$$\varepsilon^*(\mathcal{T}, P, d, D) = \inf\{\varepsilon > 0| \sup_{d(P,Q) < \varepsilon} D(T(P), T(Q)) = \infty\}.$$

*The **finite sample breakdown point (fsbp)** of $\mathcal{T}$ at a sample x_n with respect to a pseudometric D is defined by*

$$fsbp(\mathcal{T}, x_n, D) = \frac{1}{n} \min\{k \in \{1, \ldots, n\}| \sup_{Q_{n,k}} D(T(P_n), T(Q_{n,k})) = \infty\},$$

where P_n denotes the empirical distribution of x_n and $Q_{n,k}$ denotes the empirical distribution of the sample $y_{n,k}$ obtained by altering at most k points arbitrarily in x_n.

Using elements of the group equivariance structure in estimation, we now give the theorem on the upper bound for the breakdown point established by Davies and Gather (2002, 2005a).

Theorem 6.1 (Davies & Gather, 2002, 2005a) *With the above notation and under the assumption that $G_1 \neq \emptyset$, we have*

$$\varepsilon^*(\mathcal{T}, P, d, D) \leq \frac{1 - \Delta(P)}{2} \tag{6.6}$$

for all G-equivariant functionals $\mathcal{T}$, for all $P \in \mathcal{P}$ and for all pseudometrics d and D satisfying (6.1),(6.2), and (6.3).

The proof of this theorem can be found in Davies and Gather (2002, 2005a).

The upper bound given in Theorem 6.1 can also be extended to the situation of the finite sample case.

Theorem 6.2 (Davies & Gather, 2002, 2005a) *With the above notation and under the assumption $G_1 \neq \emptyset$ we have*

$$fsbp(\mathcal{T}, x_n, D) \leq \left\lfloor \frac{n - n\Delta(P_n) + 1}{2} \right\rfloor / n, \tag{6.7}$$

where P_n denotes the empirical distribution of the data x_n.

The proof can be found again in Davies and Gather (2002).

To help demonstrate the above equivariance group structure for a concrete estimation scenario, we will give an example from Davies and Gather (2002, 2005a) involving location functionals.

Example 6.1 *(Davies & Gather, 2002, 2005a) Consider p-dimensional random vectors defined by a sample space $\mathcal{X} = \mathbb{R}^p$ and the Borel $\sigma-$algebra $\mathcal{B}(\mathcal{X})$ on $\mathbb{R}^p$. We may take the family of probability measures $\mathcal{P}_\Theta$ to be the set of all distributions P on $\mathbb{R}^p$ or let $\mathcal{P}$ be the set of all $\mathbb{R}^p$-distributions with a finite expectation.*

For a distribution $P \in \mathcal{P}_\Theta$, a location functional $\mathcal{T}$ could be defined as the median of P ($\mathcal{T}(P) =$median of P) or the mean of P ($\mathcal{T}(P) = \int_{\mathbb{R}^p} x dP$). In any case, we take the parameter space to be $\Theta = \mathbb{R}^p = \{\theta : \theta \in \mathbb{R}^p\}$.

As the group acting on the sample space, we take the translation group G on $\mathbb{R}^p$ with elements $g : \mathbb{R}^p \to \mathbb{R}^p \in G$ of form $g(x) = x + a$, $a \in \mathbb{R}^p$. The group G induces a group H_G on the parameter space with elements $h_g = g$ for $g \in G$. Here the choices of D and d are not crucial and we could choose $D(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_p$ (Euclidean vector norm) and $d(P, Q) = \sup_{B \in \mathcal{C}} |P(B) - Q(B)|$, where

$$\mathcal{C} = \left\{ C : C = \{x : x^\top b + a \leq 0\} \right\}$$

with $b \in \mathbb{R}^p, a \in \mathbb{R}$. Of course, the mean and median functionals $\mathcal{T}$ are equivariant with respect to the group G.

Note that the class of constant functionals $\mathcal{T}(x) = c \in \mathbb{R}^p$ for all $x \in \mathbb{R}^p$, which naturally exhibits a breakdown point of 1, are automatically excluded from G as they fail to fulfill condition (6.4).

The group G_1 according to the definition in (6.4) is equivalent to the group G defined above except for the identity element $\{\iota\}$. As a consequence we have $\Delta(P) = P(\emptyset) = 0$ in (6.5), since no non-empty set $B \in \mathfrak{B}$ can be found such that $g_{|B} = \iota_{|B}$ holds for some $g \in G_1$.

Applying the above theorem to location functionals and the translation group yields an upper bound for the breakdown point of location functionals of $1/2$, based on $\Delta(P) = 0$.

Simply the presence of a group structure, however, does not guarantee a nontrivial upper bound for the breakdown point. If the group structure is not sufficiently rich, a trivial smallest upper bound of 1 is still possible. An example from Davies and Gather (2002) is given by the binomial model, where the data consist of a count of *successes* x based on k trials for estimating the binomial parameter $p \in [0, 1]$. This example was considered by Ruckstuhl and Welsh (2001). Davies and Gather (2002) show that there is a group acting on the sample space which consists only of two elements: the identity mapping and the mapping $g : g(x) = k - x$. But, due to the small group structure on the sample space, it turns out that the set $G_1 = \emptyset$ is empty so that Theorem 6.1 cannot be applied to find an upper bound for the breakdown point of a functional for estimating p. Indeed, equivariant functionals exist for estimating p with a breakdown point of 1 and, in fact, Davies and Gather (2002) give such a functional.

6.1.1   Dimension reduction functionals

The focus of this section is to analyze the extent to which the theory of BREAKDOWN & GROUPS can be applied to the case of dimension reduction functionals, in particular to functionals of the SIR-type. The results we will obtain however are likely to hold also for dimension reduction functionals in general. For reduction subspace estimation, the parameter space Θ corresponds to some collection of subspaces of $\mathbb{R}^p$. As a key result, we will show that, for subspace estimation, it is not possible to find a suitable pseudometric D on the parameter space Θ of $\mathbb{R}^p$-subspaces such that condition (6.3) is fulfilled. The crucial property of any D in the framework of Davies and Gather (2002, 2005a) that the pseudometric takes on the value ∞ in the supremum. As we shall discuss later, in the problem of reduction *subspace* estimation, this assumption cannot be fulfilled for *any* choice of a meaningful (pseudo)metric D on subspaces.

Elements in Breakdown and Group Structure for Sir, $\mathcal{K}$ known

To better understand BREAKDOWN & GROUPS for functionals like SIR, we now consider identifying the key elements of Davies and Gather (2002, 2005a) in the context of dimension reduction. To ease our exposition, we assume that the reduction subspace $\mathcal{B} \subset \mathbb{R}^p$ has *known* dimension $\dim(\mathcal{B}) = \mathcal{K}$ which we wish to estimate with a functional $\mathcal{T}$ associated with SIR (or similar dimension reduction functional).

PARAMETER SPACE: $\Theta_{\mathcal{K}}$.

The parameter space $\Theta_{\mathcal{K}}$ would be

$$\Theta_{\mathcal{K}} = \{\theta \subset \mathbb{R}^p : \dim(\theta) = \mathcal{K}\},$$

the set of all k-dimensional subspaces of $\mathbb{R}^p$.

STATISTICAL MODEL: $(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_{\mathcal{K}})$.

Since the data consist of observed random vectors $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, we take the sample space $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ to be defined by $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$ and $\mathfrak{B}$ as the usual Borel σ-algebra on $\mathbb{R}^p \times \mathbb{R}$. Given the form of the parameter space $\Theta_{\mathcal{K}}$ above and the nature of the SIR functional, the set of all probability measures under consideration is given by

$$\mathcal{P}_{\Theta} \equiv \mathcal{P}_{\mathcal{K}} = \{P_{(X,Y)} \text{ on } \mathbb{R}^p \times \mathbb{R} \quad | \quad \Sigma(P_{(X,Y)}) \equiv \Sigma \equiv \mathrm{Cov}(X) \text{ is positive definite}, \quad (6.8)$$
$$V(P_{(X,Y)}) \equiv V \equiv \mathrm{Cov}[\mathrm{E}(\Sigma(P_{(X,Y)})^{-1/2}X|Y)] \text{ has}$$
$$\text{ordered eigenvalues } \lambda_1 \geq \cdots \geq \lambda_{\mathcal{K}} > \lambda_{\mathcal{K}+1} \geq \cdots \geq$$
$$\lambda_p \geq 0\}.$$

Thus, $(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_{\mathcal{K}})$ describing the statistical model under consideration. The condition $\lambda_{\mathcal{K}} > \lambda_{\mathcal{K}+1}$ in (6.8) is an identifiability condition for the dimension of a subspace (parameter) estimate since we assume that $\mathcal{K}$ is known. The condition is necessary and sufficient so that, if a SIR functional $\mathcal{T}$ is applied to a distribution $P_{(X,Y)} \in \mathcal{P}_{\mathcal{K}}$, SIR will render a $\mathcal{K}$-dimensional parameter estimate $\mathcal{T}(P_{(X,Y)}) \in \Theta_{\mathcal{K}}$ consistent with the parameter space $\Theta_{\mathcal{K}}$. If SIR is applied to a distribution $P_{(X,Y)} \notin \mathcal{P}_{\mathcal{K}}$, then $\mathcal{T}(P_{(X,Y)}) \notin \Theta_{\mathcal{K}}$.

Statistical Functional: $\mathcal{T}$.

More formally, assume a functional of the SIR-type $\mathcal{T} : (\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_\mathcal{K}) \longrightarrow \Theta_\mathcal{K}$ that, when applied to $P_{(X,Y)} \in \mathcal{P}_\mathcal{K}$, yields a $\mathbb{R}^p$-subspace $\mathcal{T}(P_{(X,Y)}) = \theta = \theta(P_{(X,Y)}) \in \Theta_\mathcal{K}$ of dimension $\mathcal{K}$ given by

$$\theta = \text{span}(\Sigma^{-1/2}\eta_1, \ldots, \Sigma^{-1/2}\eta_\mathcal{K}),$$

based on the orthonormal eigenvectors $\eta_1, \ldots, \eta_\mathcal{K}$ of $V = V(P_{(X,Y)})$ corresponding to the largest $\mathcal{K}$ eigenvalues of V (which are identifiable again by (6.8)).

Group of Transformations on Sample Space: G.

A group of measurable transformations G on the sample space $\mathcal{X}$ is given by the affine transformations:

$$G = \{g : \mathbb{R}^{p+1} \to \mathbb{R}^{p+1} \mid g\binom{x}{y} \equiv \binom{g_1(x)}{g_2(y)}, \; g_2(y) = y \in \mathbb{R}, \; g_1(x) = Ax + b \in \mathbb{R}^p,$$
$$p \times p \text{ nonsingular matrix } A, b \in \mathbb{R}^p\}.$$

(6.9)

Note that transformations $g \in G$ affect only x-values, the mapping of y-values corresponds to the identity mapping. This is analogous to the regression case described in Davies and Gather (2002, 2005a), where the group acting on the sample space corresponds to the translation group and a transformation of the data is carried out only with respect to the observations of the response variable Y.

Group of Induced Transformations on Parameter Space: H_G.

The group G operating on the sample space induces a group H_G on the parameter space that corresponds to

$$H_G = \left\{h_g : \Theta \longrightarrow \Theta \mid h_g(\theta) = (A^\top)^{-1}\theta \subset \mathbb{R}^p\right\}$$

such that the equivariance property is fulfilled. (Note if $\theta \subset \mathbb{R}^p$ is $\mathcal{K}$-dimensional subspace with basis vectors, say $b_1, \ldots, b_\mathcal{K}$, then $(A^\top)^{-1}\theta$ denotes the $\mathbb{R}^p$ subspace spanned by $\text{span}((A^\top)^{-1}b_1, \ldots, (A^\top)^{-1}b_\mathcal{K})$.)

Equivariance Structure.

To prove that SIR-type functionals are equivariant w.r.t. the affine group G, we fix $P_{(X,Y)} \in \mathcal{P}_{\mathcal{K}}$ with $\mathrm{Cov}(X) = \Sigma(P_{(X,Y)}) = \Sigma$ and $V(P_{(X,Y)}) = V = \Sigma^{-1/2}W\Sigma^{-1/2}$ for $W = \mathrm{Cov}[\mathrm{E}(X|Y)]$. With a spectral decomposition, note that $V = \sum_{i=1}^{p} \lambda_i \eta_i \eta_i^\top$, where $\{\eta_i\}_{i=1}^{p}$ are the orthonormal eigenvectors of V corresponding to the ordered eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{\mathcal{K}} > \lambda_{\mathcal{K}+1} \geq \cdots \geq \lambda_p \geq 0$ by (6.8). We also have $\mathcal{T}(P_{(X,Y)}) = \theta = \mathrm{span}(\Sigma^{-1/2}\eta_1, \ldots, \Sigma^{-1/2}\eta_{\mathcal{K}})$.

Choose $g(x,y) = (Ax + b, y) \in G$. Under $g(X,Y)$ with $(X,Y) \sim P_{(X,Y)}$, or $P_{(X,Y)}^{g}$, it holds that

$$
\begin{aligned}
\overline{\Sigma} &\equiv \Sigma(P_{(X,Y)}^{g}) = A\Sigma A^\top \\
\overline{W} &\equiv W(P_{(X,Y)}^{g}) = AWA^\top \\
\overline{V} &\equiv V(P_{(X,Y)}^{g}) = \overline{\Sigma}^{-1/2}\overline{W}\overline{\Sigma}^{-1/2} = (A\Sigma A^\top)^{-1/2}(AWA^\top)(A\Sigma A^\top)^{-1/2}.
\end{aligned}
$$

We now have to find $T(P^g)$ which requires a spectral decomposition of $\overline{V}$ to find the matrix's orthogonal eigenvectors and largest $\mathcal{K}$ eigenvalues. We write $C = (A\Sigma A^\top)^{-1/2}A\Sigma^{1/2}$ and note that $A\Sigma A^\top$ and Σ are symmetric and nonnegative definite so that

$$
\begin{aligned}
CC^\top &= (A\Sigma A^\top)^{-1/2}A\Sigma A^\top(A\Sigma A^\top)^{-1/2} \\
&= \mathrm{I}_p;
\end{aligned}
$$

that is, C is a $p \times p$ orthogonal matrix. Let $\overline{\eta}_i = C\eta_i$ for $i = 1, \ldots, p$. Multiply V defined as above with C and $C^\top$ to get

$$
\begin{aligned}
\sum_{i=1}^{p} \lambda_i \overline{\eta}_i \overline{\eta}_i^\top = CVC^\top &= (A\Sigma A^\top)^{-1/2}A\Sigma^{1/2}\Sigma^{-1/2}W\Sigma^{-1/2}\Sigma^{1/2}A^\top(A\Sigma A^\top)^{-1/2} \\
&= (A\Sigma A^\top)^{-1/2}AWA^\top(A\Sigma A^\top)^{-1/2} = \overline{V}. \qquad (6.10)
\end{aligned}
$$

Since $\{\overline{\eta}_i\}_{i=1}^{p}$ are orthonormal by

$$
\overline{\eta}_i^\top \overline{\eta}_j = \eta_i^\top C^\top C \eta_j = \eta_i^\top \eta_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise}, \end{cases}
$$

we find (6.10) is the spectral decomposition of $\overline{V}$, which has eigenvectors $\overline{\eta}_1, \ldots, \overline{\eta}_\mathcal{K}$ corresponding to its well-defined, $\mathcal{K}$ largest eigenvalues (i.e, $\lambda_{\mathcal{K}+1} < \lambda_\mathcal{K}$). By definition we have

$$
\begin{aligned}
\mathcal{T}(P_{(X,Y)}^g) &= \mathrm{span}(\overline{\Sigma}^{-1/2}\overline{\eta}_1, \ldots, \overline{\Sigma}^{-1/2}\overline{\eta}_\mathcal{K}) \\
&= \mathrm{span}\left((A\Sigma A^\top)^{-1/2}C\eta_1, \ldots, (A\Sigma A^\top)^{-1/2}C\eta_\mathcal{K}\right) \\
&= \mathrm{span}\left((A\Sigma A^\top)^{-1}A\Sigma^{1/2}\eta_1, \ldots, (A\Sigma A^\top)^{-1}A\Sigma^{1/2}\eta_\mathcal{K}\right) \\
&= \mathrm{span}\left((A^\top)^{-1}\Sigma^{-1/2}\eta_1, \ldots, (A^\top)^{-1}\Sigma^{-1/2}\eta_\mathcal{K}\right) \\
&= (A^\top)^{-1}\mathrm{span}\left(\Sigma^{-1/2}\eta_1, \ldots, \Sigma^{-1/2}\eta_\mathcal{K}\right) \\
&= (A^\top)^{-1}\theta
\end{aligned}
$$

using above that $(A\Sigma A^\top)^{-1} = (A^\top)^{-1}\Sigma^{-1}A^{-1}$. Hence, the SIR-functional $\mathcal{T} : P \to \theta$ is equivariant with respect to G because

$$
\mathcal{T}(P_{(X,Y)}^g) = (A^\top)^{-1}\theta = h_g(\theta) = h_g(\mathcal{T}(P_{(X,Y)})),
$$

where $h_g(\theta) = (A^\top)^{-1}\theta \in H_G$ is the mapping induced by $g \in G$.

PSEUDOMETRIC: d.

Analogous to the location example from Davies and Gather (2002), we could choose pseudometric d defined on $\mathcal{P}_\mathcal{K}$ as $d(P_{(X,Y)}, Q_{(X,Y)}) = \sup_{B \in \mathcal{C}} |P_{(X,Y)}(B) - Q_{(X,Y)}(B)|$, where $\mathcal{C} = \left\{C : C = \{(x,y) \in \mathbb{R}^{p+1} : (x^\top, y)b + a \leq 0\}\right\}$ with $b \in \mathbb{R}^{p+1}, a \in \mathbb{R}$. This d satisfies (6.1).

PSEUDOMETRIC: D.

We now come to the point in the BREAKDOWN & GROUPS framework where there are no easy solutions. A pseudometric D on the parameter space $\Theta_\mathcal{K}$ is a mapping of the form $D : \Theta \times \Theta \longrightarrow [0, M]$ with $\sup_{\theta_1, \theta_2 \in \Theta} D(\theta_1, \theta_2) = M \in (0, \infty]$, where M denotes the supremum value of D. The problem is that any pseudometric D should satisfy two properties:

1. D should be geometrically meaningful. If $2\mathcal{K} \leq p$, then all existing metrics on subspaces from Chapter 2 would agree on the following point: two $\mathcal{K}$-subspaces $\theta_1, \theta_2 \in \Theta_{\mathcal{K}} \subset \mathbb{R}^p$ with $2\mathcal{K} \leq p$ should be maximally distant if and only if the subspaces θ_1 and θ_2 are *orthogonal*. Hence, a geometrically meaningful pseudometric D on the parameter space $\Theta_{\mathcal{K}}$ should somehow embody the notation that a maximal distance $D(\theta_1, \theta_2) = M$ between θ_1 and θ_2 means that these subspaces (of parts of them) are orthogonal. More specifically, we could say that (for any $\mathcal{K} \leq p$), if two subspaces θ_1 and θ_2 are *maximally distant* under a pseudometric D, then $\theta_1 \cap (\theta_1 \cap \theta_2)^{\perp}$ and $\theta_2 \cap (\theta_1 \cap \theta_2)^{\perp}$ should at the very least be orthogonal; see Section 2.5.2.

2. The pseudometric D must satisfy $\sup\limits_{\theta_1, \theta_2 \in \Theta} D(\theta_1, \theta_2) = M = \infty$ in (6.3) in the framework of Davies and Gather (2002, 2005a).

It is possible to find metrics which satisfy either property 1 or satisfy property 2. For example, the Frobenius metric $D(\theta_1, \theta_2) = \mathbb{F}$ fulfills the geometrical property 1 but has supremum metric value that is *finite*, namely

$$\sup_{\theta_1, \theta_2 \in \Theta_{\mathcal{K}}} \mathbb{F}(\theta_1, \theta_2) = \sqrt{2(\mathcal{K} - \max 0, 2\mathcal{K} - p)} = M < \infty$$

and so $\mathbb{F}$ does not satisfy property 2.

The sad news is that it is *mathematically impossible* to find a pseudometric D which simultaneously satisfies both properties 1 and 2 above. This can be proven as follows for the case $2\mathcal{K} \leq p$; the $2\mathcal{K} > p$ could be treated similarly. Suppose we have a pseudometric $\widetilde{D}$ which fulfills: $\sup\limits_{\theta_1, \theta_2 \in \Theta} \widetilde{D}(\theta_1, \theta_2) = \infty$ if and only if θ_1 and θ_2 become orthogonal in the $\mathbb{R}^p$. We show that such a pseudometric does not satisfy the triangle inequality (and therefore cannot be a pseudometric). Pick θ_1, θ_2 and $\theta_3 \in \mathbb{R}^p$ where the pair θ_1, θ_2 is orthogonal but the pairs θ_1, θ_3 and θ_2, θ_3 are not so that

$$\infty = \widetilde{D}(\theta_1, \theta_2) \leq \widetilde{D}(\theta_1, \theta_3) + \widetilde{D}(\theta_2, \theta_3) < \infty,$$

which is impossible.

So we have a serious dilemma. If we choose a geometrically meaningful but bounded metric, like $\mathbb{F}$, the group G_1 defined in (6.4) must be empty so that Theorem 6.1 cannot be applied to find a meaningful upper bound on the breakdown point for SIR-type functionals under the metric. This might imply that a geometrically and universally acceptable formulation of a *breakdown* for SIR-type functions is not possible. However, as shown in Chapter 4, it is still possible to use a geometrically meaningful metric like $\mathbb{F}$ to quantify some effects of contamination on subspace estimation using SIR. That is, the metric $\mathbb{F}$ allows one to mathematically study and *interpret* the effects of contamination. A solution to this dilemma may be an alternative definition for breakdown for dimension reduction functionals, and hence functionals of the SIR-type, based on an idea by Davies and Gather (2004, 2005b). For this alternative definition we refer to section 6.2.

Elements in Breakdown and Group Structure for Sir, $\mathcal{K}$ unknown

To set-up the BREAKDOWN & groups framework for SIR in the case that the dimension $\mathcal{K}$ of the reduction $\mathbb{R}^p$-subspace is unknown, we make three changes to the above structure when $\mathcal{K}$ is known. Alternative formulations of the parameter space $\Theta_{0:p}$, the family of probability distributions $\mathcal{P}_{\Theta_{0:p}}$ and the functional $\mathcal{T}$ are required.

PARAMETER SPACE: $\Theta_{0:p}$.
The parameter space $\Theta_{0:p}$ would be

$$\Theta_{0:p} = \cup_{\mathcal{K}=0}^{p}\Theta_{\mathcal{K}} = \{\theta \subset \mathbb{R}^p : \dim(\theta) = \mathcal{K},\ 0 \leq \mathcal{K} \leq p\},$$

the set of all subspaces of $\mathbb{R}^p$.

STATISTICAL MODEL: $(\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_{\Theta})$.
Again $(\mathcal{X}, \mathfrak{B}(\mathcal{X}))$ is defined as before (i.e, $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$) and

$$\mathcal{P}_{\Theta_{0:p}} \equiv \{P_{(X,Y)} \text{ on } \mathbb{R}^p \times \mathbb{R}\ \mid\ \Sigma(P_{(X,Y)}) \equiv \Sigma \equiv \text{Cov}(X)\text{ is positive definite,}$$
$$V(P_{(X,Y)}) \equiv V \equiv \text{Cov}[\text{E}(\Sigma(P_{(X,Y)})^{-1/2}X|Y)]\text{ has ordered}$$
$$\text{eigenvalues } \lambda_1 \geq \cdots \geq \lambda_{\mathcal{K}} > \lambda_{\mathcal{K}+1} \geq \cdots \geq \lambda_p \geq 0,$$
$$0 \leq \mathcal{K} \leq p\}.$$

Group Transformations: G and H_G. These remain unchanged.

Statistical Functional: $\mathcal{T}$.

A functional of the SIR-type $\mathcal{T} : (\mathcal{X}, \mathfrak{B}(\mathcal{X}), \mathcal{P}_{\Theta_{0:p}}) \longrightarrow \Theta$ that, when applied to $P_{(X,Y)} \in \mathcal{P}_{\Theta_{0:p}}$, yields a $\mathbb{R}^p$-subspace $\mathcal{T}(P_{(X,Y)}) = \theta = \theta(P_{(X,Y)}) \in \Theta$ of dimension $\mathcal{K} = \mathcal{K}(P_{(X,Y)})$ given by

$$\theta = \text{span}(\Sigma^{-1/2}\eta_1, \ldots, \Sigma^{-1/2}\eta_{\mathcal{K}}),$$

based on the orthonormal eigenvectors $\eta_1, \ldots, \eta_{\mathcal{K}}$ of $V = V(P_{(X,Y)})$ corresponding to the $\mathcal{K} = \mathcal{K}(P_{(X,Y)})$ *non-zero* eigenvalues of V.

Note that $\mathcal{T}$ is equivariant with respect to G because one can show that

$$\mathcal{T}(P_{(X,Y)}^g) = (A^\top)^{-1}\theta = h_g(\theta) = h_g(\mathcal{T}(P_{(X,Y)}))$$

as before, where $h_g(\theta) = (A^\top)^{-1}\theta \in H_G$ is mapping induced by $g \in G$. We note one interesting facet of the group induced structure H_G on Θ, where Θ consists of any subspace of $\mathbb{R}^p$. Namely, if $\theta \in \Theta$ has dimension $\dim(\theta) = \mathcal{K}$, then $h_g(\theta) \in \Theta$ also has dimension $\dim(h_g(\theta)) = \mathcal{K}$, for *any* $g \in G$. That is, under affine transformations G of the sample space, we find $\mathcal{K}$-subspaces of $\mathbb{R}^p$ must always be mapped to $\mathcal{K}$-subspaces of $\mathbb{R}^p$ for a given dimension $0 \leq \mathcal{K} \leq p$. Affine transformations of the data *cannot* alter the dimension of subspace estimates based on SIR, i.e, dimensions of $\mathcal{T}(P_{(X,Y)})$, $\mathcal{T}(P_{(X,Y)}^g)$ are equal for $g \in G$.

Pseudometrics: d and D.

The previous discussion in the $\mathcal{K}$ known case is still applicable. In particular, it is still difficult to find a geometrically meaningful metric D on Θ for which $\sup_{\theta_1, \theta_2 \in \Theta} D(\theta_1, \theta_2) = \infty$ in (6.3). An additional complication is that, for any $g \in G$ and $n \in \mathbb{N}$ and $\theta \in \Theta_{0:p}$, it holds that θ and $h_{g^n}(\theta)$ are $\mathbb{R}^p$-subspaces of the same dimension, so that it difficult to invent a pseudometric D for which $D(\theta, h_{g^n}(\theta))$ can become arbitrarily large for a given $\theta \in \Theta_{0:p}$. Worse yet, if $\theta \in \Theta_{0:p}$ has dimension $\mathcal{K} = p$ then $\theta = \mathbb{R}^p$ and we always have $h_{g^n}(\theta) = \mathbb{R}^p$, so that

$$\inf_{\theta \in \Theta_{0:p}} D(\theta, h_{g^n}(\theta)) \leq D(\mathbb{R}^p, h_{g^n}(\mathbb{R}^p)) = D(\mathbb{R}^p, \mathbb{R}^p) < \infty, \quad n \in \mathbb{N}, \ g \in G,$$

if $D(\mathbb{R}^p, \mathbb{R}^p)$ is bounded.

6.2 An alternative definition of breakdown for dimension reduction functionals ($\mathcal{K}$ known)

The discussion from the previous Section 6.1.1 indicates that, for SIR-type functionals of $\mathbb{R}^p$-subspaces of known dimension $\mathcal{K}$, a serious problem in applying the BREAKDOWN & GROUPS framework of Davies and Gather (2002, 2005a) is that a suitable pseudometric D cannot be found for the parameter space $\Theta_\mathcal{K} = \{\theta \subset \mathbb{R}^p | \dim(\theta) = \mathcal{K}\}$ of interest. Davies and Gather (2005b) suggest a different formulation of a breakdown point which is attractive because this definition does *not* even require a pseudometric D. That is, the breakdown point definition of Davies and Gather (2005b) overcomes problems suggested in the last section by completely side-stepping the need for a meaning pseudometric D to even be specified.

We next give the new breakdown point definition using the same (general) notation developed in Davies and Gather (2002, 2005a) from Section 6.1. Let $\mathcal{P}$ denote the set of all distributions on the measurable sample space $(\mathcal{X}, \mathcal{B})$. Again we suppose that there is a group G of measurable transformations of the sample space $(\mathcal{X}, \mathcal{B})$.

Davies and Gather (2005b) define a functional $\mathcal{T} : \mathcal{P}_\mathcal{T} \longrightarrow \Theta$ with $\mathcal{P}_\mathcal{T} \subset \mathcal{P}$ as equivariant with respect to the group G if the following three conditions are fulfilled

C1. $\mathcal{P}_\mathcal{T}$ is closed under all $g \in G$.

C2. $\mathcal{T}$ is well defined on $\mathcal{P}_\mathcal{T}$.

C3. $\mathcal{T}(P^g) = h_g(\mathcal{T}(P))$ for all $P \in \mathcal{P}_\mathcal{T}$ and $g \in G$.

Davies and Gather then suggest the following definition of breakdown.

Definition 6.2 *(Davies & Gather, 2005b)* *Under the above assumptions the* ***breakdown point*** *is defined as*

$$\varepsilon^*(\mathcal{T}, P, d) = \inf\{\varepsilon > 0 : d(P, Q) < \varepsilon \text{ for some } Q \notin \mathcal{P}_\mathcal{T}\},$$

where $\varepsilon^*(\mathcal{T}, P, d) = 0$ if $P \notin \mathcal{P}_\mathcal{T}$.

For SIR-Functionals when $\mathcal{K}$ is known, we have the parameter space $\Theta = \Theta_{\mathcal{K}}$ as in Section 6.1.1 with family of distributions $\mathcal{P}_{\Theta} = \mathcal{P}_{\mathcal{K}}$ from (6.8), while $\mathcal{P}$ is the set of all distributions on the sample space $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$. Now the property of identifiability of the $\mathcal{K}$ largest eigenvalues of $V(P_{(X,Y)})$ in (6.8) for $P_{(X,Y)} \in \mathcal{P}_{\mathcal{K}}$, which is necessary for the SIR functional to produce a dimension $\mathcal{K}$ subspace estimate in $\mathbb{R}^p$ (i.e., $\mathcal{T}(P_{(X,Y)}) \in \Theta_{\mathcal{K}}$), can be easily violated. For example, if a distribution $Q \in \mathcal{P}$ has a covariance matrix $V(Q)$ in (6.8) with all equal eigenvalues $\lambda_1 = \cdots = \lambda_p$ (or even just eigenvalues $\lambda_{\mathcal{K}} = \lambda_{\mathcal{K}+1}$), then $\mathcal{T}(Q)$ is not defined as SIR cannot determine a $\mathcal{K}$-dimensional subspace estimate from $V(Q)$, i.e. more equal eigenvalues than the number of dimensions $\mathcal{K}$ of the underlying e.d.r. subspace. Davies and Gather (2005b) discuss the complications of equal eigenvalues in defining a breakdown in principal component analysis and their comments are relevant for SIR.

As an example of Definition 6.2, Davies and Gather (2005b) explain that the mean is not defined for all distributions so that it follows easily that the mean has a breakdown point of 0 under the above definition for any distribution $P \in \mathcal{P}$. Since SIR depends on means (as well as covariances) of the underlying distribution, it follows that SIR must also *automatically have a breakdown point of 0* under Definition 6.2 for any $P \in \mathcal{P}$.

Upper bounds for new breakdown point with subspace functionals

Davies and Gather (2005b) also give an upper bound for the breakdown point in Definition 6.2. We can use this to find the upper bound for the breakdown point in Definition 6.2 for *any* dimension reduction functional $\mathcal{T}$ that seeks to estimate a subspace of $\mathbb{R}^p$, not just SIR-type. For this upper bound, some new notation is necessary. Define a set of distributions

$$\mathcal{P}_{ginv} = \{P \in \mathcal{P} : \text{there exists some } g \in G \text{ with } h_g \neq h_\iota, \text{ where } P^g = P\},$$

consisting of distributions in $\mathcal{P}$ which are unaffected by some transformation $g \in G$ but the corresponding parameter $\mathcal{T}(P) = \theta$ may change. According to the equivariance property of $\mathcal{T}$, we would expect for any $g \in G$ for which $P^g = P$ holds to have $\mathcal{T}(P) = \mathcal{T}(P^g) = h_g(\mathcal{T}(P))$ if $P \in \mathcal{P}_{\mathcal{T}}$, but this is of course not possible for distributions

$P \in \mathcal{P}_{ginv}$. Obviously, we can conclude that $\mathcal{P}_{ginv} \subset \mathcal{P} \setminus \mathcal{P}_{\mathcal{T}}$ for every equivariant functional $\mathcal{T}$, which implies an upper bound for the breakdown point of

$$\varepsilon^*(T, P, d) \leq \inf\{\varepsilon : d(P, Q) < \varepsilon \text{ for some } Q \in \mathcal{P}_{ginv}\} \qquad (6.11)$$

for all equivariant functionals $\mathcal{T}$ and $P \in \mathcal{P}$. Based on the above inequality, Davies and Gather (2005b) give the following upper bound for the breakdown point in Definition 6.2.

Theorem 6.3 *Suppose G contains a finite sub-group G_k of order $k \geq 2$ such that $g^k = \iota$ (the identity) holds for all $g \in G_k$. Then, it holds that*

$$\varepsilon^*(\mathcal{T}, P, d) \leq \frac{k-1}{k},$$

for any $P \in \mathcal{P}$.

In the following example, we show how to apply the above Theorem 6.3 to obtain an upper bound on the breakdown point Definition 6.2 for a subspace estimating functional. We need not assume the dimension $\mathcal{K}$ of dimension reduction is known or even that the functional corresponds to SIR. We simply suppose that the functional $\mathcal{T}$ is equivariant (i.e., satisfies *C1.-C3.*) with respect to the group of affine transformations G acting on the sample space $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$ given in (6.9). This group seems to be very appropriate for SIR-type functionals.

Example 6.2 *Suppose the sample space is $\mathcal{X} = \mathbb{R}^p \times \mathbb{R}$ with the usual Borel $\sigma-$algebra $\mathcal{B}$. Let $\mathcal{P}$ be the set of all distributions on $(\mathbb{R}^p \times \mathbb{R}, \mathcal{B})$ and take the group G of sample space transformations to be (6.9).*

Suppose $p \geq 2$. Then, for any $k \geq 2$, there is a rotation subgroup of G of order k given by

$$G_k = \left\{ g \equiv \begin{pmatrix} g_1(x) \\ g_2(y) \end{pmatrix} \in G \,\middle|\, g_1(x) = (A_k)^i x : \mathbb{R}^p \to \mathbb{R}^p, \ i \in \{1, \ldots, k\}, g_2(y) = y \right\},$$

where A_k, $k \geq 2$, is a $p \times p$ matrix given by

$$A_k = \begin{bmatrix} M_k & 0_{2 \times (p-2)} \\ 0_{(p-2) \times 2} & I_{p-2} \end{bmatrix}, \qquad M_k = \begin{bmatrix} \cos(\frac{2\pi}{k}) & -\sin(\frac{2\pi}{k}) \\ \sin(\frac{2\pi}{k}) & \cos(\frac{2\pi}{k}) \end{bmatrix}$$

and I_{p-2} *is a* $(p-2) \times (p-2)$ *identity matrix. To see this, note that*

$$(A_k)^k = \begin{bmatrix} (M_k)^k & 0_{2\times(p-2)} \\ 0_{(p-2)\times 2} & I_{p-2} \end{bmatrix} = I_p, \qquad (M_k)^k = I_2, \quad k \geq 2,$$

so that for any $k \geq 2$ *if* $g \in G_k$, *it holds that* $g_1^k(x) = (A_k)^{i \cdot k} x = \{(A_k)^k\}^i x = I_p x = x$, $x \in \mathbb{R}^p$ *with respect to some* $i \in \{1, \ldots, k\}$. *Hence, the elements* $g \in G_k$ *have order* $k \geq 2$.

If $p = 1$, *we define a subgroup of* G *of order* $k = 2$ *by*

$$G_2 = \left\{ g \equiv \begin{pmatrix} g_1(x) \\ g_2(y) \end{pmatrix} \in G \,\middle|\, g_1(x) = (-1)^i x : \mathbb{R} \to \mathbb{R}, \ i \in \{0, 1\}, g_2(y) = y \right\},$$

where clearly $g_1^2(x) = (-1)^2 x = x$, $x \in \mathbb{R}$, *if* $g \in G_2$ *when* $p = 1$.

Hence, by the existence of G_k, $k = 2$, *for any* $p \geq 1$, *it follows from Theorem 6.3 that for any dimension reduction functional* $\mathcal{T}$ *equivariant with respect to the group* G *of affine transformations (6.9), any* $P \in \mathcal{P}$, *and any pseudometric* d,

$$\varepsilon^*(\mathcal{T}, P, d) \leq \frac{1}{2}.$$

In particular, the Sir functional has a breakdown point of 0 under Definition 6.2. Davies and Gather (2005b) have argued that the median, since it is defined for all distributions, has a breakdown point of 1 under Definition 6.2.

6.3   Influence Function

The influence function (IF) was introduced by Hampel (1968, 1974) and belongs to the class of quantitative robustness measures appropriate for assessing local robustness of a functional $\mathcal{T}$. It represents a directional derivative of the functional $\mathcal{T}$ at a distribution P in the direction of a point mass in x. The influence function can be used to assess the robustness of a functional under the infinitesimal amount of contamination in a point x. A formal definition of the influence function can be found for example in Hampel et al. (1986).

Definition 6.3 (Hampel, et al. 1986) *For a real-valued statistical functional $\mathcal{T}$ at a distribution F, the **influence function** of $\mathcal{T}$ at P in a point x is given by*

$$IF(x, \mathcal{T}, P) = \lim_{\epsilon \to 0} \frac{\mathcal{T}((1-\epsilon)P + \epsilon \delta_x) - \mathcal{T}(P)}{\epsilon}, \tag{6.12}$$

where δ_x denotes the Dirac measure putting point mass 1 on x. The function $IF(x, \mathcal{T}, P)$ is defined for all x for which the above limit exists.

Hampel et al. (1986, p.41) refer to the influence function as the measure providing *the richest quantitative robustness information* as it describes the approximate and standardized effect of an additional observation (i.e., a point mass at x) on $\mathcal{T}$.

Contributions regarding the influence function in the robustness analysis of SIR can be attributed to Prendergast (2004) in his dissertation. He focused on querying the robustness of SIR towards outlying observations and violations of distributional assumptions for the data and investigated the influence function for SIR. His findings cumulated in the derivation of the influence function for functionals estimating SIR e.d.r. directions. The assumed underlying contamination model is of the form

$$P(\epsilon) = (1-\epsilon)P + \epsilon \delta_0,$$

where $\delta_0 = (x_0, y_0)$ indicates the point mass placed on the value (x_0, y_0) representing the contamination. Prendergast carried out the derivation of the influence function for a single e.d.r. direction under the following two assumptions

A.1 The slicing proportions $p_1, \ldots, p_H$ are constant, where $p_i = P(Y \in \mathbb{I}_i)$ represents the probability that the response variable Y falls into the ith slice $\mathbb{I}_i$, $i = 1, \ldots, H$.

A.2 SIR is able to recover a $\mathcal{K}$-dimensional e.d.r. subspace, if the $\mathcal{K}$ is the true dimension.

The main result by Prendergast is the derivation of the influence function the ith e.d.r. direction ($i = 1, \ldots, \mathcal{K}$) stated in Theorem 4.1 of his dissertation, which also derived for

other dimension reduction procedures, such as SAVE [cf. Cook and Weisberg (1991), Cook (2000)]. It turns out that it is not possible to present the influence function in closed form. This is because the SIR procedure/functional applied at the contaminated distribution $P(\epsilon)$ provides a certain additive component to the influence function that may not be determined without exact knowledge of the link function f.

According to Prendergast however, there exist cases for which the influence function is free of this term, such as the single index model expressed by

$$Y = \beta_1^\top X + \varepsilon,$$

with $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ a vector of regressor variables, $Y \in \mathbb{R}$ a response variable, and ε an error random variable. The corresponding mean and covariance matrix for X are denoted by μ and Σ, respectively. The sliced means $\mu_h = \mathrm{E}(X|Y \in \mathbb{I}_h) \in \mathbb{R}$ are defined for each of the H slices $\mathbb{I}_h$, $h = 1, \ldots, H$. The single index model has only one e.d.r. direction $\beta_1 \in \mathbb{R}^p$ (i.e., $\mathcal{K} = 1$) and Prendergast stated the influence function for β_1 explicitly as

$$
\begin{aligned}
IF(\beta_1, \delta_0, P) &= \left[\frac{1}{2}\beta_1\beta_1^\top - \Sigma^{-1}\right]\left[(x_0 - \mu)(x_0 - \mu)^\top - \Sigma\right]\beta_1 \qquad (6.13) \\
&\quad + \frac{1}{\lambda_1}\left(\Sigma^{-1} - \beta_1\beta_1^\top\right)\sum_{h=1}^{H} I(y_0 \in \mathbb{I}_h)\{\beta_1^\top(\mu_h - \mu)\}(x_0 - \mu),
\end{aligned}
$$

where $I(\cdot)$ denotes the indicator function and λ_1 is the largest eigenvalue of sliced version of $\mathrm{Cov}[\Sigma^{-1/2}E(X|Y)]$ given by $\sum_{h=1}^{H} p_h \Sigma^{-1/2}(\mu_h - \mu)(\mu_h - \mu)^\top \Sigma^{-1/2}$ (cf. Prendergast (2004), p. 124).

As expected the influence function is unbounded with respect to the explanatory variables and bounded with respect to the response. Based on the derivation of the influence function for the single index model Prendergast investigates in particular three types of contamination δ_0 of the regressor space, denoted by δ_1, δ_2, and δ_3, when taking $\Sigma = \mathrm{I}_p$ and $\mu = 0_p$. The term δ_1 represents a contamination in the direction of the e.d.r. direction β_1; δ_2 is contamination orthogonal to the e.d.r. direction; and a third type represents a mixture of the other two types δ_1 and δ_2. Among his findings

is that contaminations of the type δ_1 may have zero influence on the e.d.r. direction which at first glance seems to contradict our finite sample findings in Chapter 4 and the findings by Gather et al. (2002). This apparent contradiction can be explained, however. To do so, we give a new example, not presented by Prendergast.

Suppose $\Sigma = I_p$, $\mu = 0_p$ and $H = 1$ (so that $\mu_1 = \mu$), then the influence function (6.13) for the single index model becomes

$$IF(\beta_1, \delta_0, P) = \left[\frac{1}{2} \beta_1 \beta_1^\top - I_p \right] \left[x_0 x_0^\top - I_p \right] \beta_1.$$

We next assume further WLoG that $\|\beta_1\| = 1$. If the contamination term is orthogonal to β_1 (i.e., $x_0^\top \beta_1 = 0$), then the Euclidian norm of the influence function is

$$\|IF(\beta_1, \delta_2, P)\| = \frac{1}{2}.$$

Under Prendergast's contamination type δ_2, the influence function for β_1 is *bounded but not necessarily zero*. Now consider Prendergast's contamination type δ_1 using a term $x_0 = c\beta_1$ for $c \in \mathbb{R}$ so that the influence function is

$$\|IF(\beta_1, \delta_1, P)\| = \frac{1}{2} \cdot \left| c^2 - 1 \right|.$$

Under Prendergast's contamination type δ_1 with $x_0 = c\beta_1$, the influence function for β_1 is *unbounded with respect to c but could possibly be zero* if $c = 1$ or -1. This example shows that, as Prendergast suggests, contamination in the direction of β_1 (i.e, type δ_1) may not always be as bad as contamination orthogonal to β_1 (i.e., type δ_2), since the influence function is smaller when $|c| < \sqrt{2}$. But, *generally*, contamination $x_0 = c\beta_1$ in the direction of β_1 is much more influential for large values of c (i.e., $|c| > \sqrt{2}$).

Note that Prendergast assumes that Σ is unknown. When Σ is unknown but $\mathcal{K} = 1$ is known, the finite sample studies given in Chapter 4 and carried out by Hilker (1997) and Gather et al. (2002) showed that contamination $c \cdot \beta_1$ in the first e.d.r. direction can produce an estimate of β_1 that is orthogonal to β_1 when c is large. That is, under the single index model, one contamination point x_0 of large magnitude in the direction of β_1 can be dangerous for the SIR procedure. The example above, taking $x_0 = c\beta_1$ for large c, indicates that Prendergast's influence function for the single index model would seem to support this conclusion.

Chapter 7

Conclusions and Recommendations

In the focus of this thesis was a thorough investigation of the robustness properties of the dimension reduction procedure SIR (Li, (1991)). In particular, emphasis was especially paid to the finite sample behavior of the SIR procedure under data contamination.

This work builds upon the dissertation by Hilker (1997), containing the first efforts to define a breakdown point concept for dimension reduction procedures in the finite sample case, as well as research by Becker (2001) and Gather et al. (2002).

The definition of Hilker's finite sample breakdown point involved canonical correlations as a suitable distance measure between the estimated and true e.d.r. subspaces. However, as breakdown classically involves the use of an underlying metric, this definition turned out to be somewhat problematic in the sense that canonical correlations do not constitute a metric. This provided the main motivation for Chapter 2, which establishes an appropriate distance measure between $\mathbb{R}^p$-vector subspaces. Our findings cumulated in a metric based on the Frobenius matrix norm. This metric represents an adequate distance measure on $\mathbb{R}^p$-subspaces from both an intuitive and geometric point of view. The Frobenius-norm based subspace metric $\mathbb{F}$ further resolves a second drawback of Hilker's breakdown point definition. While Hilker's work stipulated that breakdown occurs if one basis vector of an estimated e.d.r. subspace is orthogonal to

the true e.d.r. subspace, we felt that it is arguably worse to estimate and select the entire orthogonal subspace of the true e.d.r. subspace of interest. The metric $\mathbb{F}$ entirely supports this notion of a worst case estimated by taking on its maximal value if and only if the orthogonal complement of the true subspace is estimated.

In Chapter 4 we considered various types of contamination which can produce a *worst case* e.d.r. subspace estimate. We demonstrated that the data contamination scenarios that produce erroneous e.d.r. subspace estimates in Sir depend for one on the knowledge of the covariance structure of the regressor variables but also on the knowledge of the dimension $\mathcal{K}$ of the true e.d.r. subspace. In particular, we showed that the type of data contamination that causes Sir to yield an erroneous subspace estimate will change depending on whether the covariance of the regressors is known or not. Based on these findings we were able to provide upper bounds for the finite sample breakdown point with respect to the metric $\mathbb{F}$ depending on the knowledge of the covariance matrix and of the dimension of the reduction subspace.

Summarizing Chapter 4, we were able to show that unlike stated at various places in the literature, Sir is indeed sensitive to outlying observations. Not only is it possible to obtain additional e.d.r. direction under data contamination as stated by Cook and Critchley (2000), e.d.r. directions can also become lost under contamination to the extent that none of the true e.d.r. directions of the e.d.r. subspace are recoverable by Sir and only the orthogonal complement of the e.d.r. subspace will ultimately be estimated.

Our theoretical findings of Chapter 4 were followed by a simulation study in Chapter 5, which clearly supported our established theory.

While our definition of the finite sample breakdown point follows geometrical intuition, it has itself the drawback of being based on a metric that can only take on finite values. This is problematic because, in *Robust Statistics*, the notion of breakdown is commonly established with respect to an unbounded metric on a parameter space (i.e., a metric that can become infinitely large when taking limits). In particular a

statistical functional is said to break down under contamination when its bias becomes infinitely large and the metric quantifying the distance between the true parameter and its contaminated estimate diverges to infinity. For $\mathbb{R}^p$-subspaces, which represent the parameters of interest in Sir, such a metric is not possible. A metric that is unbounded cannot be found as shown in Chapter 6. This difficulty with metrics is a part of our general findings in Chapter 6, where we examined the robustness of Sir at the level of population distributions based on the results of Breakdown & Groups by Davies and Gather (2005a). The main result is that Sir-Type functionals are difficult to place in the framework of Breakdown & Groups because the above mentioned problem of finding a suitable metric.

As hinted by Davies and Gather (2005b), an alternative definition of breakdown can be applied to dimension reduction functionals which may statistically be more meaningful and accurate than a definition based on common breakdown point theory, since this requires no metric on subspaces to be specified.

As a general conclusion it may be more appropriate to examine the robustness of Sir in terms of its discontinuity and lack of qualitative robustness rather than the breakdown point.

Appendix: Supporting Technical Results

Many of the proofs require basic, but perhaps less familiar, results from real analysis involving sequences. We recall some important definitions involving a sequence of real numbers $\{a_n\}_{n=1}^{\infty} \subset \mathbb{R}$ and collect some useful results with sequences in Lemma 7.1.

If $\{a_n\}_{n=1}^{\infty} \subset \mathbb{R}$ is a real sequence, define two further sequences $\{b_n\}_{n=1}^{\infty}$ and $\{c_n\}_{n=1}^{\infty}$ by $b_n = \sup\{a_n, a_{n+1}, a_{n+2}, \ldots\}$, $c_n = \inf\{a_n, a_{n+1}, a_{n+2}, \ldots\}$ for $n \in \mathbb{N}$. The sequence $\{b_n\}_{n=1}^{\infty}$ is monotonically decreasing (i.e., $b_{n+1} \leq b_n$) while $\{c_n\}_{n=1}^{\infty}$ is monotonically increasing (i.e., $c_{n+1} \geq c_n$). Define

$$\limsup_{n \to \infty} a_n = \inf\{b_n\}_{n=1}^{\infty} \qquad \liminf_{n \to \infty} a_n = \sup\{c_n\}_{n=1}^{\infty},$$

where $\limsup_{n \to \infty} a_n, \liminf_{n \to \infty} a_n \in \mathbb{R} \cup \{\infty, -\infty\}$.

Lemma 7.1 $\{a_n\}_{n=1}^{\infty} \subset \mathbb{R}$ *is a real sequence, then the following hold.*

  (i) *If $\{a_n\}_{n=1}^{\infty}$ is bounded so that $|a_n| \leq C$, $n \in \mathbb{N}$, then there exists a subsequence $\{a_{n_j}\}_{j=1}^{\infty} \subset \{a_n\}_{n=1}^{\infty}$ and real a, $|a| \leq C$, such that $a_{n_j} \to a$ as $j \to \infty$. (Bolzano-Weierstrass Theorem)*

  (ii) *If a subsequence $\{a_{n_j}\}_{j=1}^{\infty} \subset \{a_n\}_{n=1}^{\infty}$ satisfies $a_{n_j} \to a$ as $j \to \infty$ for some $a \in \mathbb{R}$, then $\liminf_{n \to \infty} a_n \leq a \leq \limsup_{n \to \infty} a_n$.*

  (iii) *$a_n \to a$ as $n \to \infty$ for some $a \in \mathbb{R}$ if and only if for any subsequence $\{a_{n_j}\}_{j=1}^{\infty} \subset \{a_n\}_{n=1}^{\infty}$, there exists a further subsequence $\{a_{n_{j_m}}\}_{m=1}^{\infty} \subset \{a_{n_j}\}_{j=1}^{\infty}$ such that $a_{n_{j_m}} \to a$ as $m \to \infty$.*

Proof. See Chapter 3 of Berberian (1994).

The following, useful result gives the singular value decomposition of a matrix A.

Lemma 7.2 *If A is a $p \times q$ matrix of rank k, then A can be expressed as*

$$A = L\Delta M^\top,$$

where L is a $p \times p$ orthogonal matrix, M is a $q \times q$ orthogonal matrix and Δ is a $p \times q$ matrix of form

$$[\mathrm{diag}(\delta_1, \delta_2, \ldots, \delta_k, \underbrace{0 \ldots, 0}_{length_{p-k}})|_{p \times (q-p)}0]$$

and $\delta_1^2, \delta_2^2, \ldots, \delta_k^2$ are the nonnegative eigenvalues of $AA^\top$ or $A^\top A$.

Proof. Kshirsagar (1972, Theorem 1, p.247) □

We use the next result to show that the product of two projection matrices must have a nonnegative trace.

Lemma 7.3 *Let $\mathcal{S}$, $\widetilde{\mathcal{S}}$ be two vector subspaces of $\mathbb{R}^p$ each spanned by an orthonormal basis $S = [s_1, \ldots, s_k]$ and $\widetilde{S} = [\widetilde{s}_1, \ldots, \widetilde{s}_{k^*}]$, respectively, where $k, k^* \leq p$. If $P_{\mathcal{S}}$ and $P_{\widetilde{\mathcal{S}}}$ denote the corresponding projection matrices onto $\mathcal{S}$ and $\widetilde{\mathcal{S}}$, respectively, then it holds that $\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}}) \geq 0$. Also, $\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}}) = 0$ if and only if the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal.*

Proof. With the projection matrices $P_{\mathcal{S}}$ and $P_{\widetilde{\mathcal{S}}}$ as defined in Definition 2.9, it follows that

$$\mathrm{tr}(P_{\mathcal{S}}P_{\widetilde{\mathcal{S}}}) = \mathrm{tr}(SS^\top \widetilde{S}\widetilde{S}^\top) = \mathrm{tr}(S^\top \widetilde{S}\widetilde{S}^\top S) = \mathrm{tr}(AA^\top),$$

where $A = S^\top \widetilde{S}$ is a $k \times k^*$ matrix. Using a singular value decomposition $S^\top \widetilde{S} = A = L\Delta M^\top$, Lemma 7.2 yields now

$$\mathrm{tr}(AA^\top) = \sum_{i=1}^{k'} \delta_i^2 \geq 0, \tag{7.1}$$

where k' represents the rank of $AA^\top$ and $0 \leq \delta_i^2 \leq 1$, $i = 1, \ldots, k'$, represent the nonnegative eigenvalues of $AA^\top$. (The property that each $\delta_i^2 \leq 1$, $i = 1, \ldots, k'$, follows from next argument. If δ_i^2 is an eigenvalue of $AA^\top$ with eigenvector v_i, $v_i^\top v_i = 1$, then

$$
\begin{aligned}
\delta_i^2 = v_i^\top v_i \delta_i = v_i^\top AA^\top v_i &= (Sv_i)^\top P_{\widetilde{S}}(Sv_i) \\
&= \|P_{\widetilde{S}} Sv_i\|^2 \\
&\leq \|Sv_i\|^2 \\
&\leq \text{largest eigenvalue of } SS^\top = P_{\mathcal{S}} \\
&= 1,
\end{aligned}
$$

using above that $P_{\widetilde{S}} = \widetilde{S}\widetilde{S}^\top$.)

Now note that if $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ are orthogonal, then $A = 0$ follows and $\text{tr}(P_S P_{\widetilde{S}}) = 0$. Conversely, if $\text{tr}(P_S P_{\widetilde{S}}) = 0$, then each eigenvalue $\delta_i^2 = 0$ in (7.1) and so $L\Delta M^\top = A = S^\top \widetilde{S} = 0_{k \times k^*}$ in the singular value decomposition. Hence, the subspaces $\mathcal{S}$ and $\widetilde{\mathcal{S}}$ have orthogonal basis vectors S, $\widetilde{S}$ and must be orthogonal. This completes the proof of Lemma 7.3. $\square$

We next summarize some useful properties of the Frobenius norm $\|\cdot\|_{\mathbb{F}}$ of a matrix, given in Definition 2.8, and relate the Frobenius norm to the spectral or matrix 2-norm $\|\cdot\|_2$ given in (2.8).

Lemma 7.4 *Let $A \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{q \times t}$ denote real-valued $p \times q$ and $q \times t$ matrices, respectively. Let $P \in \mathbb{R}^{p \times p}$ and $Q \in \mathbb{R}^{p \times p}$ denote real-valued orthogonal $p \times p$ and $q \times q$ matrices, respectively, where $PP^\top = \mathrm{I}_{p \times p}$ and $QQ^\top = \mathrm{I}_{q \times q}$. It holds that*

(a) $\|AB\|_{\mathbb{F}} \leq \|A\|_{\mathbb{F}} \cdot \|B\|_{\mathbb{F}}$

(b) $\|A\|_2 \leq \|A\|_{\mathbb{F}}$

(c) $\|A\|_{\mathbb{F}} \leq \sqrt{\min\{p, q\}} \cdot \|A\|_2$

(d) $\|PAQ\|_{\mathbb{F}} = \|A\|_{\mathbb{F}}$.

Proof. For (a), see Stewart and Sun (1990, Chapter II, Section 2.1). To establish parts (b) and (c), we use the singular value decomposition of A (see Lemma 7.2) to write

$$AA^\top = L\Delta M^\top M\Delta^\top L^\top = L\Delta\Delta^\top L^\top,$$

using the notation of Lemma 7.2 and that $M^\top M = \mathrm{I}_{q\times q}$, the $q \times q$ identity matrix, by the orthogonality of M. With this same notation, we have from (2.8) that

$$\|A\|_2^2 = \text{the largest eigenvalue of } AA^\top = \max\{\delta_1^2, \ldots, \delta_k^2\},$$

where $k = \mathrm{rank}(A) \leq \min\{p, q\}$. Note also that by $L^\top L = \mathrm{I}_{p\times p}$, it holds by definition that $\|A\|_\mathbb{F}^2$ is equal to

$$\mathrm{tr}(AA^\top) = \mathrm{tr}(L\Delta\Delta^\top L^\top) = \mathrm{tr}(\Delta\Delta^\top L^\top L) = \mathrm{tr}(\Delta\Delta^\top) = \sum_{i=1}^{k} \delta_i^2,$$

using properties of the trace operation of matrices (e.g., $\mathrm{tr}(CD) = \mathrm{tr}(DC)$ for a real-valued $p \times q$ matrix C and a real-valued $q \times p$ matrix D). Since $k \leq p$, it now follows that

$$
\begin{aligned}
\max\{\delta_1^2, \ldots, \delta_k^2\} = \|A\|_2^2 \leq \|A\|_\mathbb{F}^2 = \sum_{i=1}^{k} \delta_i^2 \ &\leq\ \min\{p, q\} \cdot \max\{\delta_1^2, \ldots, \delta_k^2\} \\
&=\ \min\{p, q\} \cdot \|A\|_2^2,
\end{aligned}
$$

establishing parts (b) and (c) of Lemma 7.4. Part (d) follows from properties of the matrix trace operation and $PP^\top = \mathrm{I}_{p\times p}$ and $QQ^\top = \mathrm{I}_{q\times q}$:

$$
\begin{aligned}
\|PAQ\|_\mathbb{F}^2 = \mathrm{tr}(PAQQ^\top A^\top P^\top) = \mathrm{tr}(PAA^\top P^\top) \ &=\ \mathrm{tr}(AA^\top P^\top P) \\
&=\ \mathrm{tr}(AA^\top) = \|A\|_\mathbb{F}^2.
\end{aligned}
$$

Note that $P^\top P = \mathrm{I}_{p\times p}$ follows above from $PP^\top = \mathrm{I}_{p\times p}$ since $P^\top$ is the unique $p \times p$ matrix inverse of the $p \times p$ matrix P. $\square$

The next result will be useful for the convergence of subspaces of $\mathbb{R}^p$, when the subspaces are spanned by a sequence of vectors. Care must be taken though because a convergent sequence of vectors does not always correspond to a convergent sequence of $\mathbb{R}^p$-subspaces. Consider the sequence of normalized vectors $b_{1,m} = (1, 0)^\top$ and $b_{2,m} =$

$(1, \frac{1}{m})^\top / \sqrt{1 + \frac{1}{m^2}}$, $m \in \mathbb{M}$, in $\mathbb{R}^2$. In this example, the space spanned by $b_{1,m}, b_{2,m}$ is $\mathbb{R}^2$ for all m but the $\mathbb{R}^2$-space spanned by $\lim_{m \to \infty} b_{1,m}, \lim_{m \to \infty} b_{2,m}$ is the line spanned by $(1,0)^\top$. Hence, limit of the span of $b_{1,m}, b_{2,m}$ (namely, $\mathbb{R}^2$) is not the span of the limit of $b_{1,m}, b_{2,m}$ (namely, $\text{span}(1,0)^\top \subset \mathbb{R}^2$). To ensure that subspaces spanned by sequences of $\mathbb{R}^p$-vectors converge as the sequence of vectors converge, we use a condition (7.2) below in Lemma 7.5.

Lemma 7.5 *Let $a \in \mathbb{R}^p$. For $k \leq p$, suppose $b_{1,m}, \ldots, b_{k,m} \in \mathbb{R}^p$, $m \in \mathbb{N}$, represent a sequence of linearly independent vectors such that, for each $i = 1, \ldots, k$*

$$\|b_{i,m}\| = 1, \ \forall\, m \in \mathbb{N}, \qquad \lim_{m \to \infty} a^\top b_{i,m} = 0,$$

and, for the $p \times k$ matrix $B_m = [b_{1,m}, \ldots, b_{k,m}]$, it holds that

$$\det(B_m^\top B_m) \geq C, \qquad m \geq N \qquad (7.2)$$

for some real $C > 0$ and $N \in \mathbb{N}$. If $\mathcal{B}_m$ represents the $\mathbb{R}^p$-subspace spanned by $\{b_{i,m}\}_{i=1}^m$ and $P_{\mathcal{B}_m}$ denotes the corresponding orthogonal projection matrix for $\mathcal{B}_m$, then

$$\lim_{m \to \infty} P_{\mathcal{B}_m} a = 0_p \qquad as \ m \to \infty.$$

Proof. Because we are considering limits, WLoG assume (7.2) holds for all $m \in \mathbb{N}$.

Because $B_m^\top B_m$ is symmetric and positive definite, we can write $B_m^\top B_m = Q_m^\top D_m Q_m$, where the $k \times k$ matrix Q_m is orthogonal (i.e., $Q_m^\top Q_m = \mathrm{I}_k$) and D_m is a diagonal matrix where the main diagonal elements consist of the positive eigenvalues $\lambda_{1,m} \geq \cdots \geq \lambda_{d,m} > 0$ of $B_m^\top B_m$; the positivity of the eigenvalues $\lambda_{i,m}$ ($i = 1, \ldots, k$, $m \in \mathbb{M}$) follows because $B_m^\top B_m$ is positive definite from $\text{rank}(B_m^\top B_m) = \text{rank}(B_m) = k$.

Since $1 = \|b_{i,m}\|^2 = b_{i,m}^\top b_{i,m}$ for $i = 1, \ldots, k$, the $k \times k$ matrix $B_m^\top B_m$ has all diagonal elements equal to 1, so that

$$\|B_m^\top\|_{\mathbb{F}}^2 = \text{tr}(B_m^\top B_m) = k, \qquad \text{tr}(B_m^\top B_m) = \text{tr}(D_m Q_m Q_m^\top) = \text{tr}(D_m) = \sum_{i=1}^{k} \lambda_{i,m};$$

it follows that $0 < \lambda_{i,m} \le k$ $(i = 1, \ldots, k,\ m \in \mathbb{M})$.

Note that $(B_m^\top B_m)^{-1} = Q_m^\top D_m^{-1} Q_m$ has ordered eigenvalues $1/\lambda_{1,m} \le \cdots \le 1/\lambda_{k,m}$ and that

$$\prod_{i=1}^{k} \frac{1}{\lambda_{i,m}} = \det\left((B_m^\top B_m)^{-1}\right) = 1/\det(B_m^\top B_m) \le \frac{1}{C}, \qquad m \in \mathbb{N}.$$

Using the last result with $1/k \le 1/\lambda_{i,m}$, it follows that

$$\left(\frac{1}{k}\right)^{k-1} \frac{1}{\lambda_{k,m}} \le \prod_{i=1}^{k} \frac{1}{\lambda_{i,m}} \le \frac{1}{C} \qquad \Rightarrow \qquad \frac{1}{\lambda_{k,m}} \le \frac{k^{k-1}}{C} \equiv \widetilde{C}$$

so that

$$
\begin{aligned}
\|(B_m^\top B_m)^{-1}\|_2^2 \;&=\; \text{largest eigenvalue of } (B_m^\top B_m)^{-1}[(B_m^\top B_m)^{-1}]^\top \\
&=\; \text{largest eigenvalue of } Q_m^\top D_m^{-2} Q_m \\
&=\; \text{largest eigenvalue of } D_m^{-2} \\
&=\; \frac{1}{\lambda_{k,m}^2} \;\le\; \widetilde{C}^2.
\end{aligned}
$$

Hence, $\|(B_m^\top B_m)^{-1}\|_{\mathbb{F}} \le \sqrt{k} \cdot \|(B_m^\top B_m)^{-1}\|_2 \le \sqrt{k} \cdot \widetilde{C}$ using Lemma 7.4. Because $\lim_{m \to \infty} a^\top B_m = 0_k^\top$ by assumption, we use Lemma 7.4 again to establish

$$
\begin{aligned}
\lim_{m \to \infty} \sqrt{(P_{\mathcal{B}_m} a)^\top (P_{\mathcal{B}_m} a)} \;&=\; \lim_{m \to \infty} \|a^\top P_{\mathcal{B}_m}\|_{\mathbb{F}} \\
&=\; \lim_{m \to \infty} \|a^\top B_m (B_m^\top B_m)^{-1} B_m^\top\|_{\mathbb{F}} \\
&\le\; \lim_{m \to \infty} \|a^\top B_m\|_{\mathbb{F}} \cdot \|(B_m^\top B_m)^{-1}\|_{\mathbb{F}} \cdot \|B_m^\top\|_{\mathbb{F}} \\
&=\; \lim_{m \to \infty} \|a^\top B_m\|_{\mathbb{F}} \cdot k \cdot \widetilde{C} \;=\; 0,
\end{aligned}
$$

from which $\lim_{m \to \infty} P_{\mathcal{B}_m} a = 0_p$ follows. $\square$

Notation

References

[1] Becker, C. (2001). Robustness Concepts for Analyzing Structured and Complex Data Sets. *Habilitationsschrift.* Department of Statistics, University of Dortmund.

[2] Bellman, R. (1961). *Adaptive Control Processes: A Guided Tour.* Princeton University Press.

[3] Berberian, S. K. (1994). *A First Course in Real Analysis.* Springer-Verlag, New York.

[4] Bond, J. C. (1999). A Robust Approach to Sir. *Ph. D. Thesis.* Department of Statistics, University of California, Berkeley.

[5] Cook, R. D. (2000). Save: A Method for Dimension Reduction and Graphics in Regression. *Commun. Statist.-Theory Meth.* **29**, 2109-2121.

[6] Cook, R. D. and Weisberg S. (1991). Comment on: Sliced Inverse Regression for Dimension Reduction. *J. Amer. Statist. Assoc.* **86**, 328-332.

[7] Cook, R. D. and Critchley, F. (2000). Identifying Regression Outliers and Mixtures Graphically. *J. Amer. Statist. Assoc.* **95**, 781-794.

[8] Crone, L. J. and Crosby, D. S. (1995). Statistical Applications of a Metric on Subspaces to Satellite Meteorology. *Technometrics.* **37**, 324-328.

[9] Davies, P. L. and Gather, U. (2002). Breakdown and Groups. *Technical Report* **57**/2002, SFB 475, University of Dortmund.

[10] Davies, P. L. and Gather, U. (2004). Breakdown and Groups II. *Technical Report* **59**/2004, SFB 475, University of Dortmund.

[11] Davies, P. L. and Gather, U. (2005a). Breakdown and Groups (with discussion and rejoinder). *Annals of Statistics.* To appear.

[12] Davies, P. L. and Gather, U. (2005b). Breakdown and Groups II. *Annals of Statistics.* To appear.

[13] Dettling, M. and Bühlmann, P. (2004). Finding Predictive Gene Groups from Microarray Data. *J. Multivariate Analysis.* **90**, 106-131.

[14] Diaconis, P. and Freedman, D. (1984). Asymptotics of Graphical Projection Pursuit. *Annals of Statistics.* **12**, 793-815.

[15] Donoho, D. L. and Huber, P. J. (1983). The Notion of Breakdown Point. In *A Festschrift for Erich Lehmann.* Edt's. P. J. Bickel, K. Doksum and J. L. Hodges, Jr., 157-184. Wadsworth, Belmont, CA.

[16] Duan, N. and Li, K.-C. (1991). Slicing regression: A link-free Regression Method. *Annals of Statistics.* **19**, 505-530.

[17] Eaton, M. (1986). A Characterization of Spherical Distributions. *Multivariate Analysis.* **20**, 272-276.

[18] Ferré, L. (1998). Determination of the Dimension in SIR and Related Methods. *J. Amer. Statist. Assoc.* **93**, 132-140

[19] Fraleigh, J. B. and Beauregard, R. A. (1990). *Linear Algebra.* Second Edition. Addison-Wesley, New York.

[20] Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *Annals of Statistics.* **19**, 1-141.

[21] Gather, U. and Becker, C. (2001). The Curse of Dimensionality – A Challenge for

Mathematical Statistics. *Jahresberichte der deutschen Mathematiker Vereinigung.* **103**, 19-36.

[22] Gather, U., Hilker, T. and Becker, C. (2001). A Robustified Version of Slice Inverse Regression, in *Statistics in Genetics and in the Environmental Sciences.* Edt's. L. T. Fernholz, S. Morgenthaler, W. Stahel. Birkhäuser, Basel, 147-157.

[23] Gather, U., Hilker, T. and Becker, C. (2002). A Note On Outlier Sensitivity of Sliced Inverse Regression. *Statistics.* **13**, 271-281.

[24] Golub, G. H. and Van Loan, C. F. (1985). *Matrix computations.* Johns Hopkins Univ. Press, Baltimore.

[25] Hall, P. and Li, K.-C. (1993). On Almost Linearity of Low Dimensional Projections from High Dimensional Data. *Annals of Statistics.* **21**, 867-889.

[26] Hampel, F. R. (1968). Contribution to the Theory of Robust Estimation. *Ph. D. Thesis.* Department of Statistics, University of California, Berkeley.

[27] Hampel, F. R. (1971). A General Qualitative Definition of Robustness. *Ann. Math. Statist.* **42**, 1887-1896.

[28] Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics – The Approach based on Influence Functions.* Wiley & Sons, New York.

[29] Harrison, D. and Rubinfeld, D. L. (1978). Hedonic Prices and the Demand for Clean Air. *J. Environ. Economics and Management.* **5**, 81-102.

[30] He, X. and Simpson, D. G. (1993). Lower Bounds for Contamination Bias: globally Minimal versus Locally Linear Estimation. *Annals of Statistics.* **21**, 314-337.

[31] Hilker, T. (1997). Robust Verfahren zur Dimensionsreduktion in Regressionsverfahren mit unbekannter Linkfunktion. *Dissertation.* Department of Statistics, University of Dortmund.

[32] Holland-Letz, T. (2001). Vergleich von Techniken zur Dimensionsbestimmung bei Verfahren vom SIR-Typ. *Master Thesis.* Department of Statistics, University of Dortmund.

[33] Huber, P. J. (1981). *Robust Statistics.* Wiley & Sons, New York.

[34] James, A. T. (1954). Normal Multivariate Analysis and the Orthogonal Group. *Annals of Mathematical Statistics.* **25**, 40-74.

[35] Kötter, T. (2000). Sliced Inverse Regression, in *Smoothing and Regression: Approaches, Computation, and Application.* Edt. Michael G. Schimek. Wiley & Sons, 497-512.

[36] Krzanowski, W . J. (1979). Between-Groups Comparison of Principal Components. *J. Amer. Statist. Assoc.* **74**, 703-707.

[37] Kshirsagar, A. M. (1972). *Multivariate Analysis.* Marcell Dekker, New York.

[38] Li, K.-C. (1991). Sliced Inverse Regression for Dimension Reduction (with discussion and rejoinder). *J. Amer. Statist. Assoc.* **86**, 316-342.

[39] 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.

[40] Li, K.-C. and Duan, N. (1987). Distribution-free and Link-free Estimation for the Sample Selection Model. *J. Econometrics.* **35**, 25-35.

[41] Lopuhaä, H. P. and Rousseeuw, P. J. (1991). Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *Annals of Statistics.* **19**, 229-248.

[42] Mathieu, M. (1998). *Funktionenanalysis: Ein Arbeitsbuch.* Spektrum Akademischer Verlag. Heidelberg, Berlin.

[43] Meyer, C. D. (2000). *Matrix Analysis and Applied Linear Algebra.* SIAM Society

for Industrial & Applied Mathematics. Philadelphia.

[44] Paige, C. C. (1984). A Note on a Result of Sun Ji-guang: Sensitivity of the CS and GSV Decomposition. *SIAM Journal of Numerical Analysis.* **21**, 186-191.

[45] Prendergast, L. (2004). Influence Functions for Sliced Inverse Regression. *Dissertation.* Department of Statistical Science, La Trobe University.

[46] Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection.* Wiley & Sons, New York.

[47] Ruckstuhl, A. F. and Welsh, A. H. (2001). Robust Fitting of the Binomial Model. *Annals of Statistics.* **29**, 1117-1136.

[48] Schott, J. R. (1994). Determining the Dimensionality if Sliced Inverse Regression. *J. Amer. Statist. Assoc.* **89**, 141-148.

[49] Schott, J. R. (1995). *Matrix Analysis for Statistics.* Wiley & Sons, New York.

[50] Stewart, G. W. and Sun, J. (1990). *Matrix Perturbation Theory.* Academic Press, Inc. San Diego.

[51] Stromberg, A. J. and Ruppert, D. (1992). Breakdown of Nonlinear Regression. *J. Amer. Statist. Assoc.* **87**, 991-997.

[52] Tibshirani, R. J., (2000). Statistical Challenges in the Analysis of DNA Microarray Data. pdf file. *Lecture delivered to National Academy of Sciences.* http://www-stat.stanford.edu/~tibs/ftp/nas.pdf.

[53] Velilla, S. (1998). Assessing the Number of Linear Components in a General Regression Problem. *J. Amer. Statist. Assoc.* **93**, 1088-1098.

[54] Xia, Y., Tong, H., Li, W. K. and Zhu, L.-X. (2002). An Adaptive Estimation of Dimension Reduction Space (with discussion and rejoinder). *J. R. Statist. Soc. B.* **64**, 363-410.