Biometrika (2017), **103**, 1, *pp*. 1–18

Printed in Great Britain

Advance Access publication on 31 July 2016

On order determination by predictor augmentation

BY WEI LUO

Center for Data Science, Zhejiang University,
866 Yuhangtang Road, Hangzhou, 310058, China
weiluo@zju.edu.cn

BING LI

Department of Statistics, The Pennsylvania State University,
326 Thomas Building, University Park, Pennsylvania 16802, USA
bing@stat.psu.edu

SUMMARY

In many dimension reduction problems in Statistics and Machine Learning, such as principal component analysis, canonical correlation analysis, independent component analysis, and sufficient dimension reduction, it is important to determine the dimension of the reduced predictor, which often amounts to estimating the rank of a matrix. This problem is called order determination. In this paper, we propose a novel and highly effective order-determination method based on the idea of predictor augmentation. We show that, if we augment the predictor by an artificially generated random vector, then the part of the eigenvectors of the matrix induced by the augmentation display a pattern that reveals information about the order to be determined. This information, when combined with the information provided by the eigenvalues of the matrix, greatly enhances the accuracy of order determination.

Some key words: Augmentation predictor; Dimension reduction; Eigenvalue; Eigenvector; Order determination.

1. INTRODUCTION

Many supervised and unsupervised statistical learning procedures operate by replacing the original predictor with a few of its linear combinations. Commonly seen examples include principal component analysis, canonical correlation analysis, independent component analysis, and sufficient dimension reduction. Most of these problems can be formulated as estimating a matrix-valued parameter M by a matrix-valued statistic $\widehat{M}$. The reduced predictors are then the projections of the original predictors on the eigenvectors of $\widehat{M}$ corresponding to its significant eigenvalues. Typically, M is a symmetric and positive semi-definite matrix.

A crucial step in all these procedures is order determination; that is, to determine the dimension of the reduced predictor. This amounts to determining the number of positive eigenvalues of M or equivalently the rank of M, based on $\widehat{M}$. As the sample estimate $\widehat{M}$ is usually of full rank, the problem becomes that of determining the set of statistically significant eigenvalues of $\widehat{M}$.

Due to the omnipresence of the order determination problem, research on this topic has been extensive. Broadly speaking, existing methods can be categorized into three types. The first type relies on the magnitude of the eigenvalues of $\widehat{M}$. This includes sequential testing procedures (Li, 1991; Li & Wang, 2007) and information criteria (Gunderson & Muirhead, 1997; Bai & Ng,

2002; Zhu et al., 2006). A sequential testing procedure gives p-values for each candidate rank, which adds to the interpretability of the results. However, as its form depends on the specific M and $\widehat{M}$ used, elaborate asymptotic expansions are often needed whenever it is applied to a new dimension reduction method. The information criteria avoid this drawback, as their forms are unified across different dimension reduction settings. However, they commonly involve tuning parameters, and appropriate choices of these parameters are highly model-dependent.

The second type of order-determination methods rely on the information contained in the eigenvectors of $\widehat{M}$. This type includes the bootstrap estimator (Ye & Weiss, 2003) and the validated information criterion (Ma & Zhang, 2015). The bootstrap estimator uses the bootstrap re-sampling to approximate the variation of the linear space spanned by the first k sample eigenvectors, and estimates the rank of M, which we denote by d, based on the tendency that the variation is large whenever $k > d$. The validated information criterion assesses whether the last $p - k$ eigenvectors of M belong to the same eigenspace, i.e. the null space of M, which holds if and only if $k \geq d$.

The third type uses information from both the eigenvalues and the eigenvectors of $\widehat{M}$, which, in many examples, pinpoints the rank of M more accurately than the previous two types. So far, the ladle estimator (Luo & Li, 2016) is the only member of this type. The idea behind the ladle estimator is to exploit the compensatory pattern between the eigenvalues and the eigenvectors. That is, when the eigenvalues are far apart, their corresponding eigenvectors tend to have small variances; but when the eigenvalues are close, their corresponding eigenvectors tend to have larger variances. The combination of both quantities yields a ladle-shaped curve whose minimum tends to occur at the true dimension d. Inspired by Ye & Weiss (2003), Luo & Li (2016) used the bootstrap to estimate the variance of the eigenvectors.

Building on the success of exploiting the information from both the eigenvalues and the eigenvectors, in this paper we introduce a different way — and in many cases a more efficient way — of extracting the information from the eigenvectors. Instead of using the bootstrap to estimate the eigenvector variation, we resort to predictor augmentation. We find that, if we add a random component to the original predictor, then the augmented part of the first d eigenvectors are smaller than the augmented part of the subsequent eigenvectors by an order of magnitude. Thus, once again, we can incorporate this information with the eigenvalues to obtain a ladle-shaped curve whose minimum tends to occur at the true dimension d. Predictor augmentation not only allows us to avoid the computationally intensive bootstrap procedure — which means we need to perform the dimension reduction many times — but also has better finite-sample performance in all the examples we considered.

To the best of our knowledge, the method proposed in this paper is the first attempt at order determination by adding noisy predictors. However, the idea of adding noisy predictors has been used previously for variable selection and screening. For example, Wu et al. (2007) introduced pseudo predictors to determine the entry significance level for forward selection by controlling two types of false selection rates. See also Luo et al. (2006), Johnson (2008), and Hu et al. (2018). Zhu et al. (2011) used pseudo predictors to find tuning parameters for nonparametric variable screening. Barber & Candès (2015, 2019) introduced knockoff variables to determine the tuning parameters in sparse regression by controlling the false discovery rate in finite sample. The key insight underlying all these methods is that we know, a priori, that the manufactured pseudo predictors bear no or little statistical relation with the response, and, if they are constructed to resemble the observed predictors, then they will give us a benchmark for what an unimportant predictor would look like statistically, thus helping us decide which predictors are important.

This insight also underlies our order-determination method: indeed, our invariance assumptions in §3 somewhat echo the knockoff condition for variable selection.

Although the invariance assumptions are weakened in §4 to a type of asymptotic contiguity, the latter assumption serves the same purpose: to ensure that the artificially created predictors are not too dissimilar to the original predictors. In this sense, the current method can be viewed as a further development of the fruitful idea of pseudo or knockoff predictors in the arena of order determination.

2. THE IDEA OF PREDICTOR AUGMENTATION

To motivate our exposition, we now use the four examples mentioned in §1 to illustrate the dimension reduction problems where order determination is needed. Let U be the random element involved. For supervised learning, U contains a predictor X, which is a vector of dimension p, and a response Y, which can be a scalar or a vector. For unsupervised learning, U only contains X itself. let $Z = \Sigma_{XX}^{-1/2}\{X - E(X)\}$ be the standardized X with zero mean and identity covariance matrix I_p, where Σ_{XX} denotes the covariance matrix of X. Since a linear combination of X can be equivalently expressed as a linear combination of Z, in some examples we use Z in place of X for easy presentation.

Example 1. In canonical correlation analysis, the goal is to find lower-dimensional linear combinations of X and Y that fully capture the linear relation between them, which amounts to estimating the leading eigenvectors of

$$M_{\mathrm{CCA}} = \Sigma_{XX}^{-1/2}\Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX}\Sigma_{XX}^{-1/2}, \tag{1}$$

where Σ_{YY} denotes the covariance matrix of Y and Σ_{XY} denotes the covariance matrix between X and Y. The estimator $\widehat{M}_{\mathrm{CCA}}$ is constructed using the sample covariance matrices. The order-determination problem is to find how many eigenvalues of $\widehat{M}_{\mathrm{CCA}}$ are significantly nonzero.

Example 2. In independent component analysis, one needs to find the linear combinations of Z that are non-normally distributed. A popular method for independent component analysis is the fourth order blind identification (Cardoso, 1989), where the non-normality is detected by excessive kurtosis characterized by the matrix

$$M_{\mathrm{ICA}} = \{E(ZZ^{\intercal}ZZ^{\intercal}) - (p+2)I_p\}^2. \tag{2}$$

Only those eigenvectors of M_{ICA} that correspond to positive eigenvalues have excess kurtosis and are of interest, and the rest are normally distributed and discarded as noise. The estimator $\widehat{M}_{\mathrm{ICA}}$ is constructed by replacing the expectation in (2) with the sample average.

Example 3. In sufficient dimension reduction, we assume that there exists a matrix $\beta \in \mathbb{R}^{p\times d}$, with $d < p$, such that

$$Y \perp\!\!\!\perp Z \mid \beta^{\intercal}Z, \tag{3}$$

where $\perp\!\!\!\perp$ means independence. For identifiable parametrization, Cook (1994, 1998) introduced the notion of the central subspace, defined as the linear space spanned by the columns of β in (3) with minimal dimension d. This space exists and is unique under fairly general conditions, and is denoted by $\mathcal{S}_{Y|X}$. Examples of sufficient dimension reduction methods include, among many others, sliced inverse regression (Li, 1991), sliced average variance estimator (Cook & Weisberg, 1991), and directional regression (Li & Wang, 2007). For a recent comprehensive account of this subject, see Li (2018).

All these methods first construct a population-level matrix M, called the candidate matrix, and estimate it by a matrix-valued statistic $\widehat{M}$. The central subspace is then estimated by the linear span of the leading eigenvectors of $\widehat{M}$. For example, the candidate matrices for sliced inverse regression and directional regression are, respectively,

$$M_{\mathrm{SIR}} = E\{E(Z \mid Y)E^{\intercal}(Z \mid Y)\}, \quad M_{\mathrm{DR}} = E[2I_p - E\{(Z - \widetilde{Z})^{\otimes 2} \mid Y, \widetilde{Y}\}]^{\otimes 2}, \tag{4}$$

where $(\widetilde{Z}, \widetilde{Y})$ is an independent copy of (Z, Y) and $A^{\otimes 2}$ denotes $AA^{\intercal}$ for any matrix A. Let $\mathcal{S}(\cdot)$ represent the column space of a matrix. Then $\mathcal{S}(M_{\mathrm{SIR}})$ is always a subspace of $\mathcal{S}(M_{\mathrm{DR}})$. If

$$E(Z \mid \beta^{\intercal} Z) \text{ is linear in } \beta^{\intercal} Z, \tag{5}$$

then $\mathcal{S}(M_{\mathrm{SIR}})$ is further a subspace of $\mathcal{S}_{Y|X}$. Similarly, under (5) and the additional condition

$$\mathrm{var}(Z \mid \beta^{\intercal} Z) \text{ is a nonrandom matrix}, \tag{6}$$

$\mathcal{S}(M_{\mathrm{DR}})$ is a subspace of $\mathcal{S}_{Y|X}$. Relaxation of (6) for directional regression can be found in Luo (2018). These conditions are exactly satisfied when Z follows a multivariate normal distribution, and hold approximately in general when p is large (Hall & Li, 1993). The order determination problem here is to estimate the rank of M_{SIR} or M_{DR}. To construct $\widehat{M}_{\mathrm{SIR}}$ and $\widehat{M}_{\mathrm{DR}}$, it is a common practice to adopt the slicing strategy, which is to partition the support of Y into H intervals; see Li (1991) and Li & Wang (2007) for details.

Example 4. In principal component analysis, we assume that $M_{\mathrm{PCA}} = \mathrm{var}(X)$ has the form

$$M_{\mathrm{PCA}} = M_0 + \sigma^2 I_p, \tag{7}$$

where M_0 is a positive semi-definite matrix and $\sigma^2 > 0$ (Jolliffe, 2002). The principal components are the eigenvectors of M_0 that span its column space, or equivalently, the eigenvectors of M_{PCA} whose corresponding eigenvalues are greater than σ^2. We use the sample covariance matrix of X as $\widehat{M}_{\mathrm{PCA}}$.

The estimation procedure in all of the above dimension reduction settings involve a sample-level matrix-valued estimator $\widehat{M}$ that converges to a population-level matrix-valued parameter M, and, with n denoting the sample size, the convergence rate of $\widehat{M}$ is typically the root-n rate.

The idea of predictor augmentation can be illustrated as follows. We augment the predictor X to X^* by an r-dimensional random vector S that is independent of U, i.e. $X^* = (X^{\intercal}, S^{\intercal})^{\intercal}$ where $S \perp\!\!\!\perp U$. For simplicity, we generate S from $N(0, I_r)$. We call X the original predictor, S the augmentation predictor, and X^* the augmented predictor. Let U^* be the corresponding augmented version of U. Apply the same estimation procedure to U^*, we have the statistic $\widehat{M}^*$, which converges to its population-level counterpart M^*. Because the column space of M^* generates the reduced predictor for U^*, and the random noise S must be absent from this reduced predictor, it is intuitive that M^* must have the form

$$M^* = \begin{pmatrix} M & 0 \\ 0 & 0 \end{pmatrix}. \tag{8}$$

In particular, M^* and M must have equal rank. This can be justified in all the dimension reduction settings mentioned above, with the exception of principal component analysis. Adjustments for principal component analysis will be discussed in §6.

LEMMA 1. *The form of M^* in (8) holds for Examples 1, 2, and 3.*

Denote the rank of M and M^* by d. Let $\lambda_1, \ldots, \lambda_{p+r}$ be the eigenvalues of M^* in descending order, that is, $\lambda_1 \geq \ldots \geq \lambda_d > 0 = \lambda_{d+1} = \ldots = \lambda_{p+r}$, and let $\beta_1, \ldots, \beta_{p+r}$ be the corresponding eigenvectors. We allow arbitrariness in β_i's if the corresponding eigenspace is multi- 165
dimensional, but assume that the eigenvectors in the same eigenspace form an orthonormal set. We define $\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{p+r}$ and $\widehat{\beta}_1, \ldots, \widehat{\beta}_{p+r}$ similarly for $\widehat{M}^*$. For any vector $v \in \mathbb{R}^{p+r}$, we call the sub-vector of its last r entries the augmentation sub-vector of v. For $k = 1, \ldots, p+r$, let $\widehat{\beta}_{k,S}$ be the augmentation sub-vector of $\widehat{\beta}_k$. Let $B_k = (\beta_1, \ldots, \beta_k)$ and $\Gamma_k = (\beta_{k+1}, \ldots, \beta_{p+r})$. By definition, Γ_d spans the null space of M^*. 170

When $k \leq d$, (8) implies that the augmentation sub-vector of β_k must be zero. The consistency of $\widehat{M}^*$ then implies the negligibility of the augmentation sub-vector $\widehat{\beta}_{k,S}$. However, this is no longer the case when $k > d$ — one can easily imagine that the length of the augmentation sub-vector $\widehat{\beta}_{k,S}$ converges to a continuous distribution because S and the noisy directions of X are intuitively exchangeable. This change of asymptotic behavior of $\widehat{\beta}_{k,S}$ before and after k reaches 175
d provides us with extra information about d, beyond that provided by the eigenvalues of $\widehat{M}^*$.

3. A MOTIVATING SPECIAL CASE

In this section, we rigorously establish the asymptotic behavior of the augmentation sub-vector of the eigenvectors of $\widehat{M}^*$ in a special case where certain invariance assumptions are satisfied. This special case is developed separately from the general case because its proof is the most 180
intuitive, and because it can be readily generalized to the high dimensional settings, as we will do in §5. We also discuss appropriate sufficient conditions for the invariance assumptions under different dimension reduction settings.

For ease of presentation, we use Z in place of X as the original predictor. Since Z and S are independent, $Z^* = (Z^\intercal, S^\intercal)^\intercal$ is the standardization of X^*. For any $(p+r-d)$-dimensional 185
orthogonal matrix A, let $M^*(A)$ and $\widehat{M}^*(A)$ be the matrix-valued parameter and its estimator, respectively, if Z^* is replaced with its rotation $(B_d, \Gamma_d A)^\intercal Z^*$. Using $(B_d, \Gamma_d A)^\intercal Z^*$ as the hypothetical augmented predictor, the reduced predictor at the population level is the projection of $(B_d, \Gamma_d A)^\intercal Z^*$ onto the column space of $M^*(A)$, or equivalently the projection of Z^* onto the column space of $(B_d, \Gamma_d A) M^*(A) (B_d, \Gamma_d A)^\intercal$. Our first invariance assumption is that the re- 190
duced predictor is invariant under the rotation of the noisy part of the augmented predictor, both at the population level and at the sample level; that is,

$$(B_d, \Gamma_d A) M^*(A) (B_d, \Gamma_d A)^\intercal = M^* \text{ and } (B_d, \Gamma_d A) \widehat{M}^*(A) (B_d, \Gamma_d A)^\intercal = \widehat{M}^*. \tag{9}$$

We call this assumption "the invariant predictor assumption". As shown next, this assumption is satisfied by all the dimension reduction methods considered in Examples 1 through 4 in §2. 195

LEMMA 2. *The invariant predictor assumption (*9*) is satisfied in all the five methods considered in Examples* 1 *through* 4*.*

Our second invariance assumption is that the distribution of the matrix-valued estimator is invariant under the rotation of the noisy part of the augmented predictor; that is,

$$\widehat{M}^*(A) \overset{\mathcal{D}}{=} \widehat{M}^*(I_{p+r-d}), \tag{10}$$ 200

where $\overset{\mathcal{D}}{=}$ means that the two random elements on both sides have the same distribution. Assumption (8) implies that $M^*(A)$ is always invariant of A; that is, $M^*(A) = M^*(I_{p+r-d})$ for all

orthogonal matrix A. The above assumption states that this property also holds for $\widehat{M}^*(A)$ in distribution. We refer to this assumption as "the invariant matrix assumption".

The invariant matrix assumption (10) requires various additional assumptions on U under different dimension reduction settings. As an example, we next illustrate such assumptions for sufficient dimension reduction. For simplicity, we assume that, in addition to $\mathcal{S}(M) \subseteq \mathcal{S}_{Y|X}$, the two spaces coincide. This means $B_d^\intercal Z^*$, which is indeed a linear combination of Z, satisfies (3). The result can be extended to the general cases by slight modifications.

LEMMA 3. *Suppose Z has a standard multivariate normal distribution and $B_d^\intercal Z^*$ can replace $\beta^\intercal Z$ in (3). Then, for any $(p+r-d)$-dimensional orthogonal matrix A,*

$$(B_d^\intercal Z^*, \Gamma_d^\intercal Z^*, Y) \overset{\mathcal{D}}{=} (B_d^\intercal Z^*, A\Gamma_d^\intercal Z^*, Y). \tag{11}$$

Clearly, (11) implies the invariant matrix assumption (10). Intuitively, the multivariate normal distribution of Z, together with the sufficient dimension reduction assumption (3), indicate that the noisy directions of Z have the standard multivariate normal distribution and are marginally independent of $(B_d^\intercal Z^*, Y)$. Since S also has the standard multivariate normal distribution and is marginally independent of $(B_d^\intercal Z^*, Y)$, the random vector $\Gamma_d^\intercal Z^*$, which consists of the noisy directions of Z and S, must also be independent of $(B_d^\intercal Z^*, Y)$. Thus, any rotation of $\Gamma_d^\intercal Z^*$ is indistinguishable from the pattern of the augmented data.

Conversely, the invariant matrix assumption implies the exchangeability between the noisy directions of Z and the components of S, which means that the noisy directions of Z must have the standard multivariate normal distribution. Because the central subspace $\mathcal{S}_{Y|X}$ is unknown in practice, this condition needs to be strengthened to that the entire Z must have the standard multivariate normal distribution. In this sense, for sufficient dimension reduction, the invariant matrix assumption is satisfied exclusively for normally distributed original predictor.

Similar to Lemma 3, one can also show that the invariant matrix assumption (10) holds for canonical correlation analysis if Z has a multivariate normal distribution and the noisy directions $\Gamma_d^\intercal Z^*$ are independent of Y; for independent component analysis, the independent components of Z must be independent of the normally distributed noise. Details are omitted.

The invariance assumptions (9) and (10) together imply that the distribution of $(B_d, \Gamma_d A)^\intercal \widehat{M}^*(B_d, \Gamma_d A)$, and hence also those of $(B_d, \Gamma_d A)^\intercal \widehat{\beta}_k$, are invariant of the orthogonal matrix A. Thus, given $\|\Gamma_d^\intercal \widehat{\beta}_k\|$, the Euclidean norm of $\Gamma_d^\intercal \widehat{\beta}_k$, $\Gamma_d^\intercal \widehat{\beta}_k$ must follow a uniform distribution on a hyper-sphere centered at the origin and with radius $\|\Gamma_d^\intercal \widehat{\beta}_k\|$. Under Assumption (8), the augmentation sub-vector $\widehat{\beta}_{k,S}$ is a linear function of $\Gamma_d^\intercal \widehat{\beta}_k$, so it must be non-negligible whenever $\|\Gamma_d^\intercal \widehat{\beta}_k\|$ is non-negligible.

Throughout the section, we assume the consistency of $\widehat{M}^*$; that is,

$$\|\widehat{M}^* - M^*\| = \mathrm{o}_P(1), \tag{12}$$

where $\|\cdot\|$ denotes the spectral norm of a matrix. The assumption is fairly general, as it holds for all the examples in §2 by the weak law of large numbers. Assumptions (8) and (12) imply that $\|\Gamma_d^\intercal \widehat{\beta}_k\|$ is negligible when $k \le d$ and non-negligible otherwise. Following the discussion in the previous paragraph, this means the negligibility of $\widehat{\beta}_{k,S}$ when $k \le d$ and non-negligibility of $\widehat{\beta}_{k,S}$ otherwise. Consequently, the function $k \mapsto \|\widehat{\beta}_{k,S}\|^2$ can characterize d as its unique jumping point. This property is formally established in the following theorem. We denote a sequence of

random variables W_n by $\Omega_P^+(1)$ if they are non-negligible in probability; that is,

$$\lim_{n\to\infty} P(W_n > \delta_n) = 1 \text{ for any } \delta_n = \mathrm{o}_P(1).$$

For more descriptions about $\Omega_P^+(1)$, see the Supplementary Material and also Luo & Li (2016).

THEOREM 1. *If Assumptions (*8*), (*9*), (*10*), and (*12*) are satisfied, then the following statements hold:*

(i) for any $k = 1, \ldots, d$, $\|\widehat{\beta}_{k,S}\|^2 = o_P(1)$;
(ii) for any $k = d+1, \ldots, p$, $\|\widehat{\beta}_{k,S}\|^2 = \Omega_P^+(1)$.

The use of the Euclidean norm in Theorem 1 is rather an arbitrary choice. Other commonly used norms, or appropriate monotonically increasing functions of these norms, can also be employed and deliver the same result.

By Theorem 1, if we draw the plot of $\|\widehat{\beta}_{k,S}\|^2$ against $k = 1, \ldots, p$, then we will observe a substantial jump from nearly zero at $k = d$ to significantly positive at $k = d+1$ when the sample size is large enough. As seen in the next section, with the aid of additional regularity conditions on $\widehat{M}^*$, this characterization of d also holds when the invariant matrix assumption (10) fails. It can then be used along with the sample eigenvalues to carry out effective order determination.

4. THE PREDICTOR AUGMENTATION ESTIMATOR

As mentioned in §2, under Assumption (8), the augmentation sub-vector is zero for any vector lying in the column space of M^*. Intuitively, the consistency assumption (12) alone is sufficient to guarantee that the augmentation sub-vectors of the leading eigenvectors of $\widehat{M}^*$ be asymptotically negligible. However, the proof of statement (ii) of Theorem 1 (see the Supplementary Material) involves both the invariant predictor assumption (9) and the invariant matrix assumption (10), the latter of which, as discussed below Lemma 3, is not granted. Hence, the proof of statement (ii) of Theorem 1 is not applicable in the general case.

For this reason, in the absence of the invariant matrix assumption (10), we make the following assumption in addition to (12). This assumption is also to replace the invariant predictor assumption (9), although the generality of the latter is justified in Lemma 2. First, we slightly generalize the concept of contiguous probability measures. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let q be a positive integer, and $\mathcal{R}^q$ be the Borel σ-field on $\mathbb{R}^q$. Let $V_n : \Omega \to \mathbb{R}^q$ be a sequence of random vectors, and $P_n = P \circ V_n^{-1}$ the distribution of V_n. We say that V_n is contiguous with respect to the Lebesgue measure μ on $(\mathcal{R}^q, \mathbb{R}^q)$, denoted by $V_n \lhd \mu$, if for any sequence of sets $A_n \in \mathcal{R}^q$, $\mu(A_n) \to 0$ implies $P_n(A_n) \to 0$. Contiguity (see, for example, Li & Babu 2019) is commonly defined with respect to two sequences of probability measures. But here, one of the sequences is taken as the Lebesgue measure. We assume

$$f(n)\mathrm{vech}\{\Gamma_d^\intercal \widehat{M}^* \Gamma_d - (\Gamma_d^\intercal \widehat{M}^* B_d)(B_d^\intercal \widehat{M}^* B_d)^{-1}(B_d^\intercal \widehat{M}^* \Gamma_d)\} \lhd \mu \tag{13}$$

where $f(n)$ is an appropriate function of n and $\mathrm{vech}(\cdot)$ stacks the columns of the upper triangle of a symmetric matrix into a vector. The transformation of $\widehat{M}^*$ in (13) can be interpreted as the squared remainder of the noisy directions $\Gamma_d^\intercal(\widehat{M}^*)^{1/2}$ after removing the effect of the informative directions $B_d^\intercal(\widehat{M}^*)^{1/2}$, much like the residual sum of squares in linear regression.

In many cases, including all the examples in §2, $\widehat{M}^*$ is a statistical functional of the empirical distribution of U^*, with the leading term of its von Mises expansion being zero, and the second

term of this expansion converging to a quadratic function of a multivariate normal distribution. Some examples of von Mises expansions for dimension reduction methods can be found in (Li, 2018, Chapter 9). Under these circumstances, Assumption (13) is satisfied for $f(n) = n$. Hence, it is much weaker than the invariant matrix assumption (10). In particular, it does not require X or S to have a multivariate normal distribution. Furthermore, since $f(n)$ is flexible, Assumption (13) can potentially cover the nonparametric dimension reduction methods such as outer product gradient and minimum average variance estimator (Xia et al., 2002).

Intuitively, Assumption (13) regulates $\widehat{M}^*$ so that, after zooming in by a factor of $f(n)$, the transformed matrix inside the brace of the left-hand side of (13) does not fall into a region of Lebesgue measure zero with non-negligible probability. Consequently, the eigenvectors of this transformed matrix will not contain asymptotically negligible entries, as the set of all $(p + r - d)$-dimensional symmetric matrices whose eigenvectors contain zero entries has Lebesgue measure zero. By Lemma A in the Supplementary Material, these eigenvectors are close to those of $\widehat{M}^*$ corresponding to non-significant eigenvalues, subject to left-multiplication of the latter by Γ_d. Hence, the augmentation sub-vector of the latter must also be non-negligible, which yields the same pattern as discussed in §3. This is formulated in the following theorem.

THEOREM 2. *If Assumptions (*8*), (*12*), and (*13*) are satisfied, then*

(i) for any $k = 1, \ldots, d$, $\|\widehat{\beta}_{k,S}\|^2 = o_P(1)$;
(ii) for any $k = d + 1, \ldots, p$, $\|\widehat{\beta}_{k,S}\|^2 = \Omega_P^+(1)$.

Theorem 2 justifies that, similar to the special case where the invariance matrix assumption (10) holds, in general the magnitude of the augmentation sub-vector of the kth sample eigenvector also stays negligible before k reaches $d + 1$, jumps to a significantly positive value at $k = d + 1$, and remains large thereafter.

Next, we use this pattern to construct an estimator of the order d. To further stabilize the pattern, we generate the augmentation predictor s times independently, and conduct dimension reduction on each augmented sample. Specifically, for $j = 1, \ldots, s$, let $\widehat{\beta}_{k,S,j}$ be the augmentation sub-vector of the kth eigenvector at the jth replication. We use the average $s^{-1} \sum_{j=1}^{s} \|\widehat{\beta}_{k,S,j}\|^2$ to characterize d. Clearly, this average also has the pattern described in Theorem 2.

We follow the same idea of Luo & Li (2016) to combine the information provided by predictor augmentation with the eigenvalues of $\widehat{M}^*$. Define the objective function $\Phi : \{0, \ldots, p\} \to \mathbb{R}$:

$$\Phi(k) = \textstyle\sum_{i=0}^{k} (s^{-1} \sum_{j=1}^{s} \|\widehat{\beta}_{i,S,j}\|^2) + \widehat{\lambda}_{k+1} / (1 + \sum_{i=1}^{k+1} \widehat{\lambda}_i), \tag{14}$$

where $\widehat{\beta}_{0,S,j}$ is set to be the origin in $\mathbb{R}^r$. We estimate d by minimizing $\Phi(\cdot)$; that is,

$$\widehat{d} = \arg\min\{\Phi(k) : k = 0, \ldots, p\}. \tag{15}$$

Because the estimator is characterized by augmenting the original predictor, we call it the predictor augmentation estimator.

In the first term of (14), we employ the accumulation of $s^{-1} \sum_{j=1}^{s} \|\widehat{\beta}_{i,S,j}\|^2$ over $i = 0, \ldots, k$, instead of the single term $s^{-1} \sum_{j=1}^{s} \|\widehat{\beta}_{k,S,j}\|^2$, to elevate Φ when $k > d$. This modification does not affect the asymptotic behavior of the objective function that uses the single summand $s^{-1} \sum_{j=1}^{s} \|\widehat{\beta}_{k,S,j}\|^2$, but it makes a subtle but important difference when d is much smaller than p and the sample size is limited: in that case, the accumulation makes the first term of (14) sufficiently large as k approaches p, so that the value of Φ at d is still less than those at large values of k, even though the $(d + 1)$th sample eigenvalue is not yet small due to limited sample size.

Another relative advantage of the modification in the high-dimensional settings will be discussed in §5.

Following the similar adjustment in the ladle estimator (Luo & Li, 2016), we introduce the denominator in the second term of (14) as a normalization, so that the objective function $\Phi(\cdot)$ is robust to the scale change of $\widehat{M}^*$. The constant one is included in the denominator as a regularization for the case $d = 0$, that is, when all the eigenvalues of $\widehat{M}^*$ are $\mathrm{o}_P(1)$. In addition, as the denominator increases with k, it sharpens the decreasing pattern of the second term of $\Phi(\cdot)$, which enhances the effectiveness of the predictor augmentation estimator when the $(d+1)$th sample eigenvalue is not small due to the limited sample size. In contrast to the adjustment in the ladle estimator that employs all the eigenvalues of $\widehat{M}^*$ at once, the denominator used here only involves the first $k+1$ eigenvalues of $\widehat{M}^*$ at each candidate rank k. This is crucial to the asymptotic study of $\Phi(k)$ for small k in the high-dimensional settings in §5.

When the matrix-valued parameter M is zero, as is the case when $d = 0$, all the sample eigenvectors have non-negligible augmentation sub-vectors, and all the sample eigenvalues are negligible, which, together, make Φ asymptotically minimized at zero. At the other extreme, when M is of full rank, all the sample eigenvectors have negligible augmentation sub-vectors, and all the sample eigenvalues are non-negligible except for the $(p+1)$th that corresponds to the augmentation predictor, which, together, make Φ consistently minimized at p. Thus the predictor augmentation estimator is still applicable in these two cases.

When M is of reduced rank but nonzero, for any $k < d$, the eigenvalue term in (14) is large; for any $k > d$, the eigenvector term in (14) is large. Moreover, since we shift the sample eigenvalues one unit to the left in the eigenvalue term in (14), both the eigenvector term and the eigenvalue term are small at $k = d$. Thus, similar to the ladle plot, the graph of Φ tends to reach its minimum at d, resulting in consistent order determination. This is proved in the following theorem.

THEOREM 3. *Suppose Assumptions (*8*), (*12*), and (*13*) are satisfied. The predictor augmentation estimator $\widehat{d}$ defined in (*15*) is consistent in the sense that*

$$\lim_{n\to\infty} P(\widehat{d} = d) = 1.$$

As mentioned earlier, the predictor augmentation estimator shares the same advantage as the ladle estimator, as it also combines the information from both the eigenvalues and the eigenvectors for order determination, which in principle is more effective than the other existing methods. Compared with the ladle estimator, the predictor augmentation estimator does not involve bootstrap re-sampling or rely on the corresponding self-similarity condition (Luo & Li, 2016), so it is computationally more efficient and potentially less demanding on the sample size.

5. CONSISTENCY IN HIGH-DIMENSIONAL SETTINGS

As mentioned in §1, many dimension reduction methods have been adapted to the high-dimensional settings, where the dimension p increases with the sample size n at some rate. For example, Zhu et al. (2006) studied the consistency of sliced inverse regression when $p = o(n^{1/2})$. Li (2007) and Chen et al. (2010) assumed that the central subspace only involves a few components of the original predictor — only a few rows and columns of M are nonzero — and modified $\widehat{M}$ by incorporating the lasso penalty, which have been shown effective in practice when $p < n$. Lin et al. (2019) incorporates the same sparsity structure on M_{SIR} and proposed the lasso-sliced inverse regression, which is consistent when $p = o(n^2)$.

In this section, we prove the consistency of the predictor augmentation estimator in the high-dimensional settings. Because the argument is greatly simplified under the invariant predictor assumption (9) and the invariant matrix assumption (10), we will make these assumptions in the high-dimensional settings. The general case is deferred to future research.

As p increases, M becomes a sequence of matrices of increasing dimensions, whose ranks d may also increase with p. Here, we assume that d is fixed for all large p. This assumption has been adopted frequently in the literature of high-dimensional sufficient dimension reduction (Zhu et al., 2006; Yu et al., 2016; Luo, 2018), and our experiences also suggest that a few linear combinations of the predictor often explain most of the variations of the response, making the rest of the predictor essentially noise.

In general, to establish any consistent order determination, it is a minimal requirement that the matrix-valued parameter M is consistently estimated; that is,

$$\|\widehat{M} - M\| = \mathrm{O}_P(\omega_{p,n}) \tag{16}$$

where $\omega_{p,n} = \mathrm{o}(1)$ as p and n diverges and $\|\cdot\|$ again denotes the spectral norm of a matrix. It is the reader's choice either to set $\omega_{p,n}$ at a general $\mathrm{o}(1)$ for simplicity, or to specify a convergence rate of $\omega_{p,n}$ for a particular dimension reduction method, e.g. $\omega_{p,n} = p^{1/2}n^{-1/2}$ for principal component analysis and sliced inverse regression (Johnstone & Lu, 2009; Lin et al., 2018). We do not impose any additional restriction on p beyond that required by (16).

Because the consistency of matrix estimation typically depends on the order of the predictor's dimension, we assume that the order of magnitude of r does not exceed that of p; that is,

$$r = \mathrm{O}(p). \tag{17}$$

Under (16) and (17), it is reasonable to assume the consistency of $\widehat{M}^*$ in terms of

$$\|\widehat{M}^* - M^*\| = \mathrm{O}_P(\omega_{p,n}) \tag{18}$$

which reduces to (12) if we set $\omega_{p,n}$ at a general $\mathrm{o}(1)$.

THEOREM 4. *Suppose Assumptions (8), (9), (10), (17), and (18) are satisfied. If, furthermore, r satisfies $rp^{-1}\omega_{p,n}^{-1} \to \infty$ as n and p diverge, then the predictor augmentation estimator $\widehat{d}$ defined in (15) is consistent in the sense that*

$$\lim_{n,p\to\infty} P(\widehat{d} = d) = 1.$$

When $\omega_{p,n}$ is set at a general $\mathrm{o}(1)$, (17) and Theorem 4 together require r to have the same order of magnitude as p. When a convergence order is specified for $\omega_{p,n}$, a smaller order of magnitude is allowed for r. The lower limit of r depends on how fast $\omega_{p,n}$ converges to zero. It is not surprising that the consistency of $\widehat{d}$ requires a certain rate of r: when r is too large, an oversized augmentation predictor will hamper the estimation accuracy of $\widehat{M}^*$; when r is too small, the augmentation predictor only has a minor contribution to the augmented predictor, rendering negligible the magnitude of the augmentation sub-vectors of all the eigenvectors of $\widehat{M}^*$, due to the unit-length restriction of the eigenvectors. In the literature of using augmentation predictor, Wu et al. (2007) empirically studied the insensitivity of their method to the choice of r, and Johnson (2008), Zhu et al. (2011), and Barber & Candès (2015, 2019) simply took $r = p$. Clearly, the characterization of the choice of r here is more specific.

In the high-dimensional settings, the accumulated form in the eigenvector term of (14) is crucial to the consistency of the predictor augmentation estimator. Had a single summand been used, the objective function Φ would still be negligible at $k = d$ and bounded below from zero

in probability at each $k > d$. However, for the function Φ to be minimized at d, we need

$$\Phi(d) < \min\{\Phi(k) : k > d\} \tag{19}$$

in probability. Considering that the cardinality of the set on the right-hand side grows to infinity with n, it would be much harder for this inequality to hold; in fact, the inequality may even fail without further conditions. The accumulated form ensures that the eigenvector term of (14) monotonically increases with k, so that, together with the positive semi-definiteness of $\widehat{M}^*$, inequality (19) will be satisfied if a single summand $s^{-1}\sum_{j=1}^{s}\|\widehat{\beta}_{d+1,S,j}\|^2$ in the eigenvector term is greater than the $(d+1)$th sample eigenvalue, which is much easier.

To our knowledge, only the information criteria have been shown to be consistent in the literature for order determination in the high-dimensional settings (Bai & Ng, 2002; Zhu et al., 2006). However, the tuning parameters for these methods depend heavily on the order of dimension p and the convergence rate of the matrix estimator, which makes it hard to find tuning constants that work universally well across different models and dimensions. By contrast, the proposed estimator has the same simple form without additional adjustment in the high-dimensional settings, making it more stable across different model and dimension settings.

Though we have made the invariant matrix assumption (10) in Theorem 4, we can relax it in the following way. As mentioned at the beginning of this section, consistency of the estimation of M in the high-dimensional settings is often achieved by making a certain sparsity assumption on M, accompanied by a sparse estimate $\widehat{M}$. In this case, the task of order determination is to estimate the rank of the nonzero (symmetric) sub-matrix of M, and we can apply our predictor augmentation estimator to the corresponding sub-matrix of $\widehat{M}$. If it is reasonable to assume that the nonzero sub-matrix of M has a bounded dimension as p grows, then the consistency of $\widehat{d}$ is governed by Theorem 3, which does not require the invariant matrix assumption (10).

6. Adjustment for principal component analysis

Principal component analysis has been commonly used as a simple and effective means of dimension reduction for many decades. As (7) indicates, order determination for this method amounts to estimating the smallest number of leading eigenvalues of M_{PCA} such that all the remaining eigenvalues are equal to a positive constant σ^2. Thus, it differs slightly from the previously discussed dimension reduction settings, where the remaining eigenvalues are zero.

As mentioned in §2, Assumption (8) is not feasible for principal component analysis, as it would require the augmentation predictor to be degenerate. Because M_{PCA} is of full rank, this assumption is not useful either: if it were true then the augmentation predictor would distinguish itself from the noisy directions of the original predictor, defeating the purpose of augmentation. Ideally, an augmentation predictor should have covariance matrix $\sigma^2 I_r$, so that it is indistinguishable from the noisy directions of the original predictor, i.e.

$$M^*_{\mathrm{PCA}} = \begin{pmatrix} M_{\mathrm{PCA}} & 0 \\ 0 & \sigma^2 I_r \end{pmatrix} = \begin{pmatrix} M_0 & 0 \\ 0 & 0 \end{pmatrix} + \sigma^2 I_{p+r}. \tag{20}$$

Since σ^2 is unknown, we must estimate it from $\widehat{M}_{\mathrm{PCA}}$. For this purpose, we assume

$$d < p/2, \tag{21}$$

which is mild in practice, as a reasonably small number of principal components can often explain a large percentage of variation in the data. A similar assumption can be found in Luo et al. (2009) for directional regression. Under (21), the median of the sample eigenvalues from $\widehat{M}_{\mathrm{PCA}}$,

denoted by $\widehat{\sigma}^2$, converges to σ^2 asymptotically. We generate the augmentation predictor under $N(0, \widehat{\sigma}^2 I_r)$. Again, following the discussion under (13), this choice is not unique.

We first study the consistency of the predictor augmentation estimator for principal component analysis when p is fixed. By the strong law of large numbers, Assumption (12) holds. Because $M^*_{\rm PCA}$ is nonsingular, we can apply the central limit theorem and strengthen (13) to

$$n^{1/2}\mathrm{vech}(\Gamma_d^\intercal \widehat{M}^*_{\rm PCA}\Gamma_d - \sigma^2 I_{p+r-d}) \xrightarrow{D} \mathrm{Q} \tag{22}$$

where Q is a multivariate normal distribution with nonsingular covariance matrix. Following the proof of Theorem 2 in the Supplementary Material, we can show the same pattern for the augmentation sub-vector of the eigenvectors derived from the augmented sample. On the other hand, although all the sample eigenvalues are significantly positive, they still display a decreasing pattern with a sudden drop at $k = d$ and negligible differences after $k > d$. Therefore, the objective function Φ is still consistently minimized at d. This intuition is rigorously formulated in the following theorem. The proof closely resembles that of Theorem 3, and is omitted.

THEOREM 5. *Suppose Model (7) for principal component analysis holds, and Assumptions (12), (21), and (22) are satisfied. Then the predictor augmentation estimator $\widehat{d}$ defined in (15) is consistent in the sense that*

$$\lim_{n\to\infty} P(\widehat{d} = d) = 1.$$

In the literature, principal component analysis has also been studied under the high-dimensional settings (Zou et al., 2006; Johnstone & Lu, 2009). Following the reasoning in §5, we expect the predictor augmentation estimator to be consistent under these settings as well, subject to the normality of the original predictor and a more careful estimation of the baseline eigenvalue σ^2, etc. The detailed development will be left to future research.

Finally, it is conceivable that the adjustment presented in this section also applies to the similar order determination problems where one needs to detect the number of largest eigenvalues of M when the rest are equal but nonzero.

7. SIMULATION STUDIES

We now use simulated models to investigate the effectiveness of the predictor augmentation estimator. Since the estimator can be applied to various dimension reduction settings and may favor normally distributed predictors, for comprehensiveness, we investigate it under each dimension reduction setting mentioned in §2, and under various distributions of the original predictor.

For principal component analysis, we use

$$\text{Model 1: } X = \Sigma_{XX}^{1/2} Z,$$

where $\Sigma_{XX}^{1/2}$ is a diagonal matrix with diagonal elements $(2, 2, 2, 0.5, \ldots, 0.5)$, and Z follows a uniform distribution on the hypersphere $\{z \in \mathbb{R}^p : \|z\|^2 = p\}$. The number of leading eigenvalues of the matrix-valued parameter $M_{\rm PCA}$ in (7) is $d = 3$.

For canonical correlation analysis, we let Y be a p-dimensional random vector with

$$Y_1 = X_1 + X_2 + \varepsilon_1,\ Y_i = X_{i+1} + \varepsilon_i,\ \text{for } i = 2, \ldots, d,\ Y_j = 2\,\varepsilon_j\ \text{ for } j > d, \tag{23}$$

where X has a standard multivariate t distribution with five degrees of freedom and $\varepsilon_1, \ldots, \varepsilon_p$ are independent errors distributed as $N(0, 0.5^2)$. The rank of $M_{\rm CCA}$ in (1) is d. We first let $d = 2$, and label (23) by Model 2. To evaluate the performance of the proposed estimator when d is larger, we also set d to be the smallest integer to the right of $p^{1/2}$, and label (23) by Model 2^*.

For independent component analysis, we use:

$$\text{Model 3: } X = AU,$$

where A is a square matrix whose diagonal entries are one and off-diagonal entries are 0.5, and U consists of independent components, with the first two components having the exponential distribution with mean equal to one and the other components having the standard normal distribution. The rank of M_{ICA} in (2) is $d = 2$.

For sufficient dimension reduction, we use

$$\begin{aligned}
&\text{Model 4: } Y = \sin(X_1) + \varepsilon,\\
&\text{Model 5: } Y = X_1 + X_2^2 + \varepsilon,\\
&\text{Model 6: } Y = X_1^2 + X_2^2 + \varepsilon,
\end{aligned}$$

where $\varepsilon \perp\!\!\!\perp X$ and $\varepsilon \sim N(0, 0.5^2)$. For Model 4, X follows the same distribution as Z in Model 1; for Models 5 and 6, X follows the standard multivariate normal distribution.

Since condition (5) is satisfied for Models 4, 5, and 6, we apply sliced inverse regression to all these models. As the method cannot detect variations in Y that are symmetric about X, the ranks of M_{SIR} in these models are $d = 1$, $d = 1$, and $d = 0$, respectively. Since condition (6) is satisfied in Models 5 and 6, we apply directional regression to these two models. Now with the symmetric patterns recoverable, the rank of M_{DR} is $d = 2$ for both models.

We compare our predictor augmentation estimator (PA) with four types of order-determination methods. The first type is the various sequential testing procedures designed for different dimension reduction settings. For canonical correlation analysis, we apply $\mathrm{S_F}(\mathrm{T_d^2})$ introduced by Fujikoshi (1977). For both sliced inverse regression and directional regression, we follow Bura & Yang (2011) to use the weighted chi-square test and the Wald-type chi-square test, and refer to them as BY1 and BY2, respectively. We choose the numbers of slices to be $H = 10$ and $H = 3$ for implementing $\widehat{M}_{\mathrm{SIR}}$ and $\widehat{M}_{\mathrm{DR}}$, respectively. The significance level for all the sequential testing procedures is taken to be $\alpha = 0.05$. For principal component analysis and independent component analysis, we are not aware of any sequential testing procedures available for order determination.

The second type of methods are those based on various forms of information criteria. Specifically, we use the $\mathrm{PC_{p1}}$ criterion (Bai & Ng, 2002) for principal component analysis, the $\widehat{\mathrm{K}}_{\mathrm{MC}}$ criterion (Gunderson & Muirhead, 1997) for independent component analysis, and the Bayesian information criterion (Zhu et al., 2006) for both sliced inverse regression and directional regression. In implementing Zhu et al.'s method, we fix the number of slices H at $n/20$ as they suggested and use their suggested tuning parameter.

As mentioned in the Introduction, the above two types of methods are based on the eigenvalues of $\widehat{M}$. The third method is the aforementioned validated information criterion (Ma & Zhang, 2015) based on the eigenvector variations of $\widehat{M}$, and the fourth method is the aforementioned ladle estimator (Luo & Li, 2016) based on both the eigenvalues and eigenvector variations of $\widehat{M}$. Both of these methods apply to all the four dimension reduction settings we considered. To make the situation clear, we list the various comparison scenarios in Table 1.

From an omitted simulation experiment, we found that the performance of the predictor augmentation estimator is not sensitive to the repetition number s, so we set it to ten. To evaluate the effect of the dimension r of S on the performance of the predictor augmentation estimator, we choose r to be the smallest integer to the right of $p/20$, $p/10$, $p/5$, $p/2$, respectively, and additionally choose r to be p. The corresponding estimators are denoted by PA1, PA2, PA3, PA4,

Table 1: Index for order-determination methods

Method	Model	d	ST	IC	VIC	Ladle	PA
PCA	1	3	–	$\mathrm{PC}_{\mathrm{p1}}$	VIC	Ladle	PA
CCA	2	2	$\mathrm{S}_{\mathrm{F}}(\mathrm{T}_{\mathrm{d}}^2)$	$\widehat{\mathrm{K}}_{\mathrm{MC}}$	VIC	Ladle	PA
CCA	2^*	$\lceil p^{1/2} \rceil$	$\mathrm{S}_{\mathrm{F}}(\mathrm{T}_{\mathrm{d}}^2)$	$\widehat{\mathrm{K}}_{\mathrm{MC}}$	VIC	Ladle	PA
ICA	3	2	–	–	VIC	Ladle	PA
SIR	4	1	(BY1, BY2)	ZMP	VIC	Ladle	PA
SIR	5	1	(BY1, BY2)	ZMP	VIC	Ladle	PA
SIR	6	0	(BY1, BY2)	ZMP	VIC	Ladle	PA
DR	5	2	(BY1, BY2)	ZMP	VIC	Ladle	PA
DR	6	2	(BY1, BY2)	ZMP	VIC	Ladle	PA

"PCA" stands for the principal component analysis, "CCA" for the canonical correlation analysis, "ICA" for the independent component analysis, "SIR" for sliced inverse regression, and "DR" for directional regression. "ST" stands for the sequential testing procedures, where "BY1" and "BY2" denote the weighted chi-square test and the Wald-type chi-square test, respectively, proposed by Bura & Yang (2011), "IC" for information criteria, where "ZMP" denotes the Bayesian information criterion for sufficient dimension reduction introduced by Zhu et al. (2006), "VIC" for validated information criterion, and "PA" for predictor augmentation estimator. $\lceil p^{1/2} \rceil$ denotes the smallest integer to the right of $p^{1/2}$.

Table 2: Comparison of order-determination methods at $p = 10$

Method	Model	n	d	ST	IC	VIC	Ladle	PA1	PA2	PA3	PA4	PA5
PCA	1	50	3	–	30	100	99	100	100	100	100	100
CCA	2	100	2	93	97	100	92	89	99	99	100	100
CCA	2^*	100	4	93	96	53	99	98	100	100	100	100
ICA	3	500	2	–	–	0	90	86	95	95	87	69
SIR	4	200	1	(89, 0)	0	100	97	91	99	99	100	100
SIR	5	200	1	(89, 22)	0	34	89	86	98	100	100	99
SIR	6	200	0	(88, 3)	100	100	25	70	91	98	99	100
DR	5	200	2	(78, 91)	0	6	81	98	99	98	90	0
DR	6	200	2	(64, 96)	0	9	55	85	96	99	98	90

Entries in Columns $5 - 13$ are the percentages of correct order determination for the corresponding method and model, based on 1000 runs.

and PA5. In addition to the standard normal distribution, we also generate S in the same way as Z in Model 1. The results are similar and deferred to the Supplementary Material.

We first take $p = 10$, which means $d = 4$ in Model 2^* and $r = 1, 2, 3, 6, 10$, respectively, in the five predictor augmentation estimators. To make the matrix estimator $\widehat{M}$ reasonably accurate but not too accurate to make order determination trivial, we let $n = 50$ for principal component analysis, $n = 100$ for canonical correlation analysis, $n = 500$ for independent component analysis, and $n = 200$ for sufficient dimension reduction. For each model and each order-determination method, we generate 1000 simulation samples, and record the percentage of correct order determination. The results are presented in Table 2.

Table 2 shows that the effectiveness of the sequential testing procedures heavily depends on the models they are applied to, especially for the Wald-type test, which consistently misspecifies d for sliced inverse regression in Models 4 and 6, but correctly specifies d for directional regression in Model 6. Overall, the sequential testing procedures fail to reach their nominal significance

Table 3: Comparison of order-determination methods at $p = 80$

Method	Model	n	d	ST	IC	VIC	Ladle	PA1	PA2	PA3	PA4	PA5
PCA	1	100	3	–	87	100	100	100	100	100	100	100
CCA	2	400	2	96	100	0	100	95	100	100	100	100
CCA	2^*	400	9	93	100	0	100	100	100	100	0	0
SIR	4	400	1	$(0, 0)$	0	0	99	93	99	100	100	93
SIR	5	400	1	$(0, 0)$	0	0	85	91	99	100	96	14
SIR	6	400	0	$(100, 0)$	0	100	27	51	92	99	100	100

level 0.05 at the current sample sizes. The performance of the information criteria, in particular the Bayesian information criterion, also varies considerably from model to model, indicating that the tuning parameter suggested in Zhu et al. (2006) may not be universally optimal. The validated information criterion, while mostly superior to the ordinary information criteria, still performs rather poorly for four models.

By contrast, both the ladle estimator and the five predictor augmentation estimators clearly outperform the others. The former fails only in Model 6. Among the five predictor augmentation estimators, the choice of $r = 3$ is optimal, with the choices of $r = 2$ and $r = 6$ being generally comparable. All these three choices lead to consistently high performance of the resulting estimator across all the models, superior also to the ladle estimator. The choice of $r = 1$ also leads to a consistent predictor augmentation estimator but is clearly suboptimal to the above three choices, and the choice of $r = 10$ leads to an estimator that is mostly consistent but completely fails when applied to directional regression in Model 5. These observations reconfirm the theoretical trade-off in choosing r discussed in §5.

We then increase the dimension p to 80. We take $n = 100$ for principal component analysis, and $n = 400$ for both canonical correlation analysis and sliced inverse regression. These represent the cases where p is moderately large compared with n. Independent component analysis and directional regression are omitted because, at this dimension, they require much larger sample sizes to be effective. Since the invariant matrix assumption is satisfied only by Models 5 and 6, the asymptotic theory in §5 is appropriate only for these two models. At the current dimension p, d equals nine in Model 2^*, and r in the five predictor augmentation estimators are 5, 9, 17, 41, and 80, respectively. We generate 1000 simulation samples, and record the percentage of correct order determination for each method in Table 3.

A comparison of Table 3 with Table 2 shows that the adverse effect of increased dimension is severe on the sequential testing procedures, the information criteria, and the validated information criterion, which consistently misspecify d in most models. By contrast, this effect is negligible on the ladle estimator and the predictor augmentation estimators. The performance of the ladle estimator is actually improved, becoming comparable with even the optimal predictor augmentation estimator for Models 1, 2, 2^*, and 4. This is likely due to the enlarged sample size that enhances the effectiveness of bootstrap re-sampling. Similar to Table 2, the choice of r as the smallest integer to the right of $p/5$ is optimal for the proposed estimator, followed closely by r being the smallest integer to the right of $p/10$. Both choices are the clear winners among all the methods. The proposed estimators for $r = 5$, 41, and 80 also perform well for most models.

A comparison on the performance of the proposed estimator between Model 2 and Model 2^* indicates that a small r is preferred when d is large. This is reasonable: while the accumulation procedure in the first term of $\Phi(\cdot)$ in (14) may lift up $\Phi(d)$ to be non-negligible, a small r down-weighs each summand of this term and balances out the adverse effect.

Table 4: Comparison of order-determination methods when $p > n$

Method	Model	n	p	d	ST	IC	VIC	Ladle	PA1	PA2	PA3	PA4	PA5
Sparse PCA	1	100	400	3	–	100	0	25	99	100	100	99	88
Lasso-SIR	7	300	600	1	–	0	0	65	100	100	100	100	100

"Sparse PCA" stands for the sparse principal component analysis proposed by Zou et al. (2006), and "Lasso-SIR" stands for the method proposed by Lin et al. (2019). The other abbreviations follow those in Table 1.

We next further increase the dimension p to be greater than n. To ensure the consistency of $\widehat{M}$, we apply the sparse principal component analysis (Zou et al., 2006) to Model 1 with $(n, p) = (100, 400)$. For sufficient dimension reduction, we follow Lin et al. (2019) to use

$$\text{Model 7: } Y = 0.5(\beta^{\intercal} X)^3 + \varepsilon$$

with $(n, p) = (300, 600)$, where both X and the first 20 entries of β are generated independently from the standard multivariate normal distribution for each simulated sample, and the rest of β are zero. We apply their lasso-sliced inverse regression to this model, which has $d = 1$. The results are summarized in Table 4. The sequential testing procedures are inapplicable, as the asymptotic distributions of the sample eigenvalues are unknown.

Table 4 shows that the predictor augmentation estimator outperforms the other methods for all the five choices of r. Together with Table 3, these results also show that the proposed estimator is consistent in the high-dimensional settings even if the invariant matrix assumption (10) is violated.

8. APPLICATION

We now use the residential building data set in Rafiei & Adeli (2015) to illustrate the effectiveness of the proposed estimator as applied with directional regression. The data set was collected to study the effect of certain physical, economic, and financial variables on the construction cost and the sale price of residential buildings, which helped to gauge the future profit before a new construction was started. 103 variables, including eight physical and financial variables and 95 economic variables falling into five different time lags, were recorded as the predictor. With one extreme outlier removed, the data contained 371 observations.

We focus on modeling the construction cost of the residential buildings using directional regression. With $r = 21$, i.e. the smallest integer to the right of $p/5$, the predictor augmentation estimator suggests that the central subspace is of dimension two. To assess the plausibility of this estimate, we conduct directional regression with $d = 3$, leading to a reduced predictor with three components. We first plot the response variable against the first component in the left panel of Figure 1, which shows a clear monotone pattern by the loess curve and the corresponding confidence band. We next plot the residual of the loess regression from the left panel against the second component in the middle panel of Figure 1, which again shows a significant pattern. Thus, a sufficient reduced predictor must be at least two-dimensional. The heteroscedasticity in both plots also suggests that the reduced predictor contributes to the conditional variance of the response variable. We then plot the residual of the nonparametric regression on the first two components against the third component, as shown in the right panel of Figure 1, which no longer shows any pattern. In particular, the confidence band of the loess fit covers a horizontal line (representing zero) entirely. Hence, the third component is redundant for modeling the response variable in the presence of the first two, indicating the minimal sufficiency of the first two components.

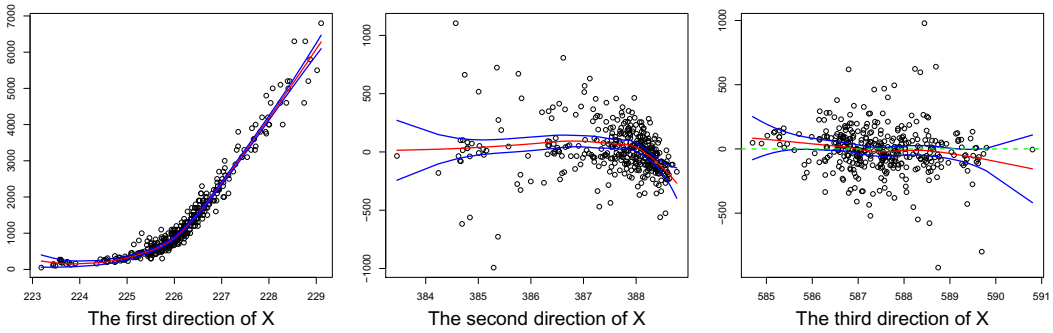

Fig. 1: In the left panel, the y-axis represents the response variable, the construction cost of residential buildings; in the middle panel, it represents the residual after the loess regression on the first component of the reduced predictor; in the right panel, it represents the residual after the loess regression on the first two components of the reduced predictor. In each panel, the solid curves are the loess regression fit and its confidence band. The horizontal dashed line in the right panel is the x-axis.

ACKNOWLEDGEMENTS

The authors would like to thank the two referees and an Associate Editor for their very helpful comments and suggestions. The research of Bing Li is supported in part by a U.S. National Science Foundation grant.

SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online contains the proofs of Lemmas 1 – 3 and Theorems 1 – 4, and properties of $\Omega_P^+(1)$ and lemmas that are useful for these proofs.

REFERENCES

BAI, J. & NG, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

BARBER, R. F. & CANDÈS, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics* **43**, 2055–2085.

BARBER, R. F. & CANDÈS, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics* **47**, 2504–2537.

BURA, E. & YANG, J. (2011). Dimension estimation in sufficient dimension reduction: a unifying approach. *Journal of Multivariate Analysis* **102**, 130–142.

CARDOSO, J.-F. (1989). Blind identification of independent components with higher-order statistics. In *Higher-Order Spectral Analysis, 1989. Workshop on*. IEEE.

CHEN, X., ZOU, C. & COOK, R. (2010). Coordinate-independent sparse suffcient dimension reduction and variable selection. *The Annals of Statistics* **6**, 3696–3723.

COOK, R. D. (1994). Using dimension reduction subspaces to identify important imputs in models of physical systems. *In 1994 Proceedings of the Section on Physical and Engineering Sciences: American Statistical Association, Alexandria, VA.,* , 18–25.

COOK, R. D. (1998). *Regression Graphics*. Wiley, New York.

COOK, R. D. & WEISBERG, S. (1991). Discussion of "Sliced inverse regression for dimension reduction". *Journal of the American Statistical Association* **86**, 316–342.

FUJIKOSHI, Y. (1977). *Asymptotic expansion for the distributions of some multivariate tests. In: Krishnaiah, P. R., ed. Multivariate Analysis. Vol. IV.* Wiley, Amsterdam: North-Holland, pp. 55–71.

GUNDERSON, B. & MUIRHEAD, R. (1997). On estimating the dimensionality in canonical correlation analysis. *Journal of Multivariate Analysis* **62**, 121–136.

HALL, P. & LI, K.-C. (1993). On almost linearity of low dimensional projections from high dimensional data. *The Annals of Statistics* **47**, 867–889.

HU, W., LABER, E. & STEFANSKI, L. (2018). Variable selection using pseudo-variables. *arXiv preprint arXiv:1804.01201* .

JOHNSON, B. A. (2008). Variable selection in semiparametric linear regression with censored data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 351–370.

JOHNSTONE, I. M. & LU, A. Y. (2009). On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association* **104**, 682–693.

JOLLIFFE, I. T. (2002). *Principal Component Analysis, Second Edition*. Springer, New York.

LI, B. (2018). *Sufficient Dimension Reduction: Methods and Applications with R*. CRC Press.

LI, B. & BABU, G. J. (2019). *A Graduate Course on Statistical Inference*. Springer, New York.

LI, B. & WANG, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association* **102**, 997–1008.

LI, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association* **86**, 316–342.

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

LIN, Q., ZHAO, Z. & LIU, J. S. (2018). On consistency and sparsity for sliced inverse regression in high dimensions. *The Annals of Statistics* **46**, 580–610.

LIN, Q., ZHAO, Z. & LIU, J. S. (2019). Sparse sliced inverse regression via lasso. *Journal of the American Statistical Association* **114**, 1726–1739.

LUO, R., WANG, H. & TSAI, C.-L. (2009). Contour projected dimension reduction. *The Annals of Statistics* **37**, 3743–3778.

LUO, W. (2018). On the second-order inverse regression methods for a general type of elliptical predictors. *Statistica Sinica* **28**, 1415–1436.

LUO, W. & LI, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika* **103**, 875–887.

LUO, X., STEFANSKI, L. A. & BOOS, D. D. (2006). Tuning variable selection procedures by adding noise. *Technometrics* **48**, 165–175.

MA, Y. & ZHANG, X. (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika* **102**, 409–420.

RAFIEI, M. H. & ADELI, H. (2015). A novel machine learning model for estimation of sale prices of real estate units. *Journal of Construction Engineering and Management* **142**, 04015066.

WU, Y., BOOS, D. D. & STEFANSKI, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association* **102**, 235–243.

XIA, Y., TONG, H., LI, W. K. & ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64**, 363–410.

YE, Z. & WEISS, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association* **98**, 968–979.

YU, Z., DONG, Y. & ZHU, L.-X. (2016). Trace pursuit: A general framework for model-free variable selection. *Journal of the American Statistical Association* **111**, 813–821.

ZHU, L., MIAO, B. & PENG, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association* **101**, 630–642.

ZHU, L.-P., LI, L., LI, R. & ZHU, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475.

ZOU, H., HASTIE, T. & TIBSHIRANI, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265–286.