Journal of Multivariate Analysis 115 (2013) 84–107

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Optimal transformation: A new approach for covering the central subspace

François Portier*, Bernard Delyon

IRMAR, University of Rennes 1, Campus de Beaulieu, 35042 Rennes Cedex, France

ARTICLE INFO

Article history:
Received 18 November 2011
Available online 23 September 2012

AMS 2000 subject classifications:
primary 62G08
secondary 62H11
62H05

Keywords:
Inverse regression
Slicing estimation
Sufficient dimension reduction
Central subspace

ABSTRACT

This paper studies a general family of methods for sufficient dimension reduction (SDR) called the test function (TF), based on the introduction of a nonlinear transformation of the response. By considering order 1 and 2 conditional moments of the predictors given the response, we distinguish two classes of methods. The optimal members of each class are calculated with respect to the asymptotic mean squared error between the central subspace (CS) and its estimate. Moreover the theoretical background of TF is developed under weaker conditions than the existing methods. Accordingly, simulations confirm that the resulting methods are highly accurate.

1. Introduction

Dimension reduction in regression aims at improving poor convergence rates derived from the nonparametric estimation of the regression function in large dimension. It attempts to provide methods that challenge the curse of dimensionality by reducing the number of predictors. A specific dimension reduction framework, called the *sufficient dimension reduction* has drawn attention in the last few years. Let Y be a random variable and X a p-dimensional random vector. To reduce the number of predictors, it is proposed to replace X by a number smaller than p of linear combinations of the predictors. The new covariate vector has the form PX, where P can be chosen as an orthogonal projector on a subspace E of $\mathbb{R}^p$. Clearly, this kind of methods relies on an alchemy between the dimension of E, which needs to be as small as possible, and the conservation of the information carried by X about Y through the projection on E. In the SDR literature, mainly two kind of spaces have been studied. First a *dimension reduction subspace* (DRS) [16] is defined by the conditional independence property

$$Y \perp\!\!\!\perp X \mid P_c X, \tag{1}$$

where P_c is the orthogonal projector on a DRS. In words, it means that knowing $P_c X$, there is no more information carried by X about Y. It is possible to show that (1) is equivalent to

$$\mathbb{P}(Y \in A \mid X) = \mathbb{P}(Y \in A \mid P_c X) \quad \text{a.s.}, \tag{2}$$

for any measurable set A, or there exists a noise e and a function f such that Y has the representation

$$Y = f(P_c X, e) \quad \text{with } e \perp\!\!\!\perp X.$$

* Corresponding author.
E-mail addresses: francois.portier@univ-rennes1.fr (F. Portier), bernard.delyon@univ-rennes1.fr (B. Delyon).

0047-259X/$ – see front matter
doi:10.1016/j.jmva.2012.09.001

Moreover under some additional conditions (see for instance [7]), the intersection of all the DRS is itself a DRS. Consequently, there exists a unique DRS with minimum dimension and we call it the *central subspace* (CS). In this article the CS is noted E_c. Secondly, another space called a *mean dimension reduction subspace* (MDRS) has been defined in [8] with the property

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|P_m X] \quad \text{a.s.}, \tag{3}$$

where P_m is the orthogonal projector on a MDRS. Clearly, the existence of a MDRS requires a weaker assumption than the existence of a DRS and therefore it seems to be more appropriate to the context of regression. Because of the equivalent formulation of Eq. (3),

$$Y \perp\!\!\!\perp \mathbb{E}[Y|X] \mid P_m X,$$

the definition of a MDRS imposes that all the dependence between Y and its regression function on X is carried by $P_m X$. If the intersection of all the MDRS is itself a MDRS, then it is called the *central mean subspace* (CMS) [8]. In the following the CMS is noted E_m. Finally, notice that because a DRS is a MDRS, the CS contains the CMS.

There exist many methods for estimating the CS and the CMS and these methods can be divided into two groups, those who require some assumptions on the distribution of the covariates and those who do not. The second group includes *structure adaptive method* (SAM) [15], *minimum average variance estimation* (MAVE) [21], *structural adaptation via maximum minimization* (SAMM) [6]. Those methods are free from conditions on the predictors but require a nonparametric estimation of the regression function $\mathbb{E}[Y|X = x]$. More recently, the *central solution space* (CSS) [18] has also been introduced to alleviate some common assumptions on the distribution of the predictors. In this article we are concerned only with methods of the first group. They are presented in the following paragraph.

For the sake of clarity, from now on we work in terms of standardized covariate $Z = \Sigma^{-\frac{1}{2}}(X - \mathbb{E}[X])$ with $\Sigma = \mathrm{var}(X)$ is a full rank matrix. Hence we define the standardized CS as $\Sigma^{\frac{1}{2}} E_c$. Since there is no ambiguity, we still note it E_c and we still denote by P_c the orthogonal projector on this subspace. Define d as the dimension of E_c. For any matrix M, we note $\mathrm{span}(M)$ the space generated by the column vectors of M, and $\mathrm{vec}(M)$ the vector of columns of M. The usual Kronecker product will be noted $\otimes$ and we denote by $Z^{(k)}$ the k-th component of the vector Z.

All the methods of the first group derive from the principle of inverse regression: instead of studying the regression curve which implies high dimensional estimation problems, the study is based on the inverse regression curve $\mathbb{E}[Z|Y = y]$ or the inverse variance curve $\mathrm{var}(Z|Y = y)$. We will respectively refer to the orders 1 and 2 approaches. To infer about the CS, order 1 methods require that

Assumption 1 (*Linearity Condition*)**.**

$$Q_c \mathbb{E}[Z|P_c Z] = 0 \quad \text{a.s.},$$

where $Q_c = I - P_c$. Under the linearity condition and the existence of the CS, it follows that $\mathbb{E}[Z|Y] \in E_c$ a.s. and then if we divide the range of Y into H slices $I(h)$, we have for every h,

$$m_h = \mathbb{E}[Z|Y \in I(h)] \in E_c, \tag{4}$$

and clearly, the space generated by some estimators of the m_h's estimates the CS, or more precisely a subspace of the CS. To obtain a basis of this subspace, [16] proposed a principal component analysis and this led to an eigendecomposition of the matrix

$$\widetilde{M}_{\mathrm{SIR}} = \sum_h p_h m_h m_h^T, \tag{5}$$

where $p_h = \mathbb{P}(Y \in I(h))$. Many methods relying on the inverse regression curve such as *sliced inverse regression* (SIR) [16] have been developed. Other ways to estimate the inverse regression curve are investigated in *kernel inverse regression* (KIR) [24] and *parametric inverse regression* (PIR) [2]. Instead of a principal component analysis, the minimization of a discrepancy function is studied in *inverse regression estimator* (IRE) [9] to obtain a basis of the CS. In [23], some polynomial transformations of the response are considered to estimate some subspaces of the CS. For a complete background about order 1 methods, we refer to [9].

By considering regression models like $Y = |Z^{(1)}| + e$, with Z having a symmetric distribution and $e \perp\!\!\!\perp Z$, some authors (for instance [16]) noticed that sometimes, $\mathbb{E}[Z|Y] = 0$ a.s. and refer to the SIR pathology when it occurs. Order 2 methods have been introduced to handle such a situation. In addition to the linearity condition order 2 methods require that

Assumption 2 (*Constant Conditional Variance (CCV)*)**.**

$$\mathrm{var}(Z|P_c Z) = Q_c \quad \text{a.s.},$$

then under the linearity condition, CCV and the existence of the CS, it follows that $\mathrm{span}(\mathrm{var}(Z|Y) - I) \in E_c$ a.s. and by considering a slicing of the response, we have

$$\mathrm{span}(v_h - I) \subset E_c, \tag{6}$$

where $v_h = \mathrm{var}(Z|Y \in I(h))$. Since the spaces generated by the matrices $(v_h - I)$'s are included in the CS, *sliced average variance estimation* (SAVE) in [10] proposed to make an eigendecomposition of the matrix

$$\widetilde{M}_{\mathrm{SAVE}} = \sum_h p_h (v_h - I)^2,$$

to derive a basis of the CS. Another combination of matrices based on the inverse variance curve is *sliced inverse regression-II* (SIR-II) [16]. More recently, *contour regression* (CR) [20], and *directional regression* (DR) [19] investigate a new kind of estimator based on empirical directions. Besides, methods for estimating the CMS also require Assumptions 1 and 2. They include *principal Hessian direction* (pHd) [17], and *iterative Hessian transformation* (IHT) [8]. In order to clear the failure of certain methods when facing pathological models and to keep their efficiency in other cases, some combinations of the previous methods as SIR and SIR-II, SIR and pHd or SIR and SAVE have been studied in [14,22].

As we have just highlighted, Assumptions 1 and 2 are needed to respectively characterize the CS with the inverse regression curve and the inverse variance curve. A first point is that the linearity condition and CCV assumed together are really close to an assumption of normality on the predictors. Moreover for each quoted method, these assumptions guarantee only that the estimated CS is asymptotically included in the true CS. A crucial point in SDR and a recent new challenge is to propose some methods that allow a comprehensive estimation of the CS under mild conditions. Recent researches are concerned with this problem, [20,19] proposed a new set of assumptions that guarantees the exhaustiveness of the estimation, i.e. the whole CS is estimated.

In this paper, we propose a general point of view about SDR by introducing the test function method (TF). The original basic idea of TF is to investigate the dependence between Z and Y by introducing nonlinear transformations of Y, and inferring about the CS through their covariances with Z or ZZ^T. Actually, an important difference between TF and other methods is that neither the inverse regression curve and nor the inverse variance curve are estimated as it is suggested by Eqs. (4) and (6). In this paper, these two curves are some working tools but the inference about the CS is obtained through some covariances. More precisely, the CS is obtained either by an inspection of the range of

$$\mathbb{E}[Z\psi(Y)],$$

when ψ varies in a well chosen finite family of function or either by an eigendecomposition of

$$\mathbb{E}[ZZ^T\psi(Y)],$$

where ψ is a well chosen function. Hence two kinds of methods can be distinguished, the order 1 test function methods (TF1) and the order 2 test function methods (TF2). Notice that $\widetilde{M}_{\mathrm{SIR}}$ is an estimate of $\mathbb{E}[Z\mathbb{E}[Z|Y]^T]$, hence SIR may be seen as a particular case of TF1. For similar reasons, TF1 also extends results of [23] who considered polynomial transformations. Besides, the results regarding TF2 are somewhat more interesting because just a single transformation $\psi : \mathbb{R} \to \mathbb{R}$ is sufficient to have an accurate estimate. As a consequence, there are few connections between TF2 and the order 2 existing methods, for instance SAVE and DR involve transformations of the form $\mathbb{E}[ZZ^TA(Y)]$ where $A(Y)$ is a matrix.

This paper has two principal objectives: to provide a general theoretical study of TF1 and TF2 linked with the background of the existing methods, and to derive the optimal members of each methodology through an asymptotic variance minimization. The optimal members are respectively called order 1 optimal function (OF1) and order 2 optimal function (OF2), they correspond to two distinguish methods for the estimation of the CS. As a result, a significant improvement in accuracy is targeted by OF1 and OF2. We show that TF allows to relax some hypotheses commonly assumed in the literature, especially we alleviate the CCV hypothesis for TF2. Moreover for both methodology TF1 and TF2, we provide mild conditions ensuring an exhaustive characterization of the CS. The present work is divided into the three following principal parts:

- Existence of the CS and the CMS
- Exhaustiveness of TF
- Optimality for TF.

More precisely, it is organized as follows. In Section 2, we propose some new conditions ensuring the existence of the CS and the CMS. Section 3 is devoted to TF1: we present some conditions that guarantee the exhaustiveness of the method and then we calculate the optimal transformation of the response for TF1 to minimize the estimation error. By following the same path, we study TF2 in Section 4. Accordingly, we propose two plug-in methods derived from the minimization of the mean squared error: OF1 and OF2. The estimation of the dimension of the CS is addressed in Section 5. Finally, in Section 6 we compare both methods to existing ones through simulations.

2. Existence of the central subspace and the central mean subspace

Conditions on the uniqueness of subspaces that allow a dimension reduction are investigated in this section. This problem has drawn attention early in the literature but it seems not to be the case any more. As a consequence of the definition of the CS (resp. CMS), its existence is equivalent to the uniqueness of a DRS (resp. MDRS) with minimal dimension. In [7, Proposition 6.4 p. 108], it is shown that the existence of the CS can be obtained by constraining the distribution of X to have a convex density support. Moreover, in [8], the existence of the CMS is ensured under the same condition than the CS. We prove in Theorem 1 below that the convexity assumption can be significantly weakened.

Theorem 1. *Under (1), if X has a density such that the Lebesgue measure of the boundary of its support is equal to 0, then the CS and the CMS exist.*

The proof is postponed to Appendix A. Since TF is only concerned about the CS estimation, we assume from now on its existence.

3. Order 1 test function

A way to introduce TF1 is to consider some relevant facts about the SIR estimation. As explained in the Introduction, SIR consists of estimating the matrix

$$M_{\mathrm{SIR}} = \mathbb{E}\left[Z\mathbb{E}[Z|Y]^T\right],$$

whose column space is included in the CS. To make that possible, a slicing approximation of the conditional expectation $\mathbb{E}[Z|Y]$ is conducted and it leads to $\widetilde{M}_{\mathrm{SIR}}$ of Eq. (5). Because $p_h > 0$, it is clear that

$$\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = \mathrm{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}], h = 1, \ldots, H\right),$$

and it follows that SIR estimates a subspace spanned by the covariances between Z and a family of Y-measurable functions. The first goal of TF1 is to extend SIR to some other families of functions Ψ_H, in order to estimate E_c with span $(\mathbb{E}[Z\psi(Y)], \psi \in \Psi_H)$. Besides, notice that

$$\widetilde{M}_{\mathrm{SIR}} = \mathbb{E}\left[Z\left(\phi_1(Y), \ldots, \phi_p(Y)\right)\right],$$

with $\phi_k(y) = \sum_h \alpha_{k,h}\mathbb{1}_{\{y\in I(h)\}}$ and $\alpha_{k,h} = \mathbb{E}[Z^{(k)}|Y \in I(h)]$. It follows that

$$\mathrm{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}], h = 1, \ldots, H\right) = \mathrm{span}\left(\mathbb{E}[Z\phi_k(Y)], k = 1, \ldots, p\right),$$

and clearly SIR synthesizes the information contained in a set of H vectors into a set of p vectors. Although each of these spaces are equal, it is not the case for their respective estimators with finite sample. Accordingly, another issue for TF1 is to choose the p functions ϕ_k's in order to minimize the variance of the estimation.

The following theorem is not new at all. Yet, it makes a simple link between TF1 and the CS. We introduce the function space $L_p(r(\omega))$ defined as

$$L_p(r(\omega)) = \{\psi : \mathbb{R} \to \mathbb{R}; \mathbb{E}[|\psi(Y)|^p r(\omega)] < +\infty\},$$

where $r : \mathbb{R} \to \mathbb{R}_+$ is a measurable function and ω a random variable.

Theorem 2. *Assume that Z satisfies Assumption 1 and has a finite first moment. Then, for every measurable function $\psi \in L_1(\|Z\|)$, we have*

$$\mathbb{E}[Z\psi(Y)] \in E_c.$$

The linearity condition is often equated with an assumption of sphericity on the distribution of the predictors. It is well known that if Z is spherical then it satisfies the linearity condition but the converse is false. Actually, linearity condition and sphericity are not so closely related: in [12], it is shown that a random variable Z is spherical if and only if $\mathbb{E}[QZ|PZ] = 0$ for every rank 1 projector P and $Q = I - P$. Clearly, at this stage, the sphericity seems to be a too large restriction to obtain the linearity condition. However unlike the sphericity, since we do not know P_c, the linearity condition could not be checked on the data. For instance, an assumption close to the linearity condition is to ask the distribution of Z to be invariant by the orthogonal symmetry to the space E_c, i.e. $Z \overset{d}{=} (2P_c - I)Z$. Then for any measurable function f,

$$\mathbb{E}[Q_c Z f(P_c Z)] = -\mathbb{E}[Q_c Z f(P_c Z)],$$

which implies the linearity condition. Recalling that sphericity means invariance in distribution by every orthogonal transformation, we have just shown that an invariance in distribution by a particular one suffices to get the linearity condition. Moreover, the assumption of sphericity suffers from the fact that if we add to Z some independent components, the resulting vector is no longer spherical whereas the linearity condition is still satisfied.

3.1. Exhaustiveness for TF1

As a consequence of Theorem 2, spaces generated by $(\mathbb{E}[Z\psi_1], \ldots, \mathbb{E}[Z\psi_H])$ are included in E_c. Our goal is to obtain the converse inclusion. Because TF1 is an extension of SIR, this one has a central place in the following argumentation. We start by giving a necessary and sufficient condition for covering the entire CS with SIR. Then under the same condition we extend SIR to a new class of methods.

Assumption 3 (*Order 1 Coverage Condition*)**.** For every nonzero vector $\eta \in E_c$, $\mathbb{E}[\eta^T Z|Y]$ has a nonzero variance.

The previous assumption is clearly equivalent to $\mathrm{span}(M_{\mathrm{SIR}}) = E_c$. Moreover, it is always true that for H large enough $\mathrm{span}(M_{\mathrm{SIR}}) = \mathrm{span}(\widetilde{M}_{\mathrm{SIR}})$. Then we have the equivalent form

$$\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = E_c$$

which was called the coverage condition in [9]. Nevertheless we use the former to make a link with some assumptions developed in [19] (see below Assumption 5 for more details). The aim is to shed light on some coverage-type result replacing the conditional expectation $\mathbb{E}[Z|Y]$ in M_{SIR} by some known and finite family of functions. Particularly, the previous equation provides such a result but only for the family of indicator functions.

Theorem 3. *Assume that Z and Y satisfy Assumptions 1 and 3. Assume also that Z has a finite second moment. If Ψ is a total countable family in the space $L_1(\|Z\|)$, then one can extract a finite subset Ψ_H of Ψ such that*

$$\mathrm{span}\left(\mathbb{E}[Z\psi(Y)],\ \psi \in \Psi_H\right) = E_c.$$

Remark 1. According to Theorem B.2, quoted in Appendix B, we can apply Theorem 3 with any family of functions that separates the points, for example polynomials, complex exponentials or indicators. Especially for polynomials, we extend a result stated in [23, Proposition 4], whose purpose is that E_c can be covered with the family $\Psi_H = \{Y^h,\ h = 1, \ldots, H\}$ if H goes to infinity.

To make possible a simple use of this theorem we need to recall this result. If $u = (u_1, \ldots, u_H)$ is a family of vectors in $\mathbb{R}^p$, then $\mathrm{span}(uu^T) = \mathrm{span}(u)$. Thus, if we denote by $\psi_1, \ldots, \psi_H$ some elements of a family that separates the points, then the CS can be obtained by making an eigendecomposition of the order 1 test function matrix associated with the functions $\psi_1, \ldots, \psi_H$ defined as

$$M_{\mathrm{TF1}} = \sum_{h=1}^{H} \mathbb{E}[Z\psi_h(Y)]\mathbb{E}[Z\psi_h(Y)]^T.$$

Especially, under the conditions of Theorem 3, the eigenvectors associated with a nonzero eigenvalue of M_{TF1} generate E_c. Moreover, as pointed out before, for H large enough $\mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = \mathrm{span}(M_{\mathrm{SIR}})$. A proof of this result is cleared up by Theorem 3. By applying it with the family of indicator functions, it gives that

$$\mathrm{span}\left(\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h = 1, \ldots, H\right) = \mathrm{span}(\widetilde{M}_{\mathrm{SIR}}) = \mathrm{span}(M_{\mathrm{SIR}}) = E_c,$$

for H is sufficiently large. Moreover, SIR can be understood as a particular TF1. Expression (5) implies that

$$\widetilde{M}_{\mathrm{SIR}} = \sum_{h=1}^{H} p_h^{-1}\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}]\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}]^T,$$

then, SIR is equivalent to TF1 realized with the weighted family of indicator functions $\left(\frac{\mathbb{1}_{\{Y\in I(h)\}}}{\sqrt{p_h}}\right)$. Besides, for any family of functions, the space spanned by M_{TF1} is invariant by positive weighting of the functions. Nevertheless with a finite sample, it is no longer the case for the estimated space and intuitively it seems that such a weighting could influence the convergence rate and improve the accuracy of TF1. The choice of the weights for the family of indicators is debated is Section 3.2.

3.2. Optimality for TF1: OF1

In this section, we develop a plug-in method based on the minimization of the variance estimation in the case of the family of indicator functions for Ψ_H. Theorem 3 and Remark 1 imply that the whole subspace E_c can be covered by the family of vectors $\{\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],\ h = 1, \ldots, H\}$ for a suitable partition $I(h)$. To provide a basis of E_c, it suffices to extract d orthogonal vectors living in this space. This procedure is realized by SIR. Nevertheless, the issue here is somewhat more complicated, we want to find d orthogonal vectors that have the smallest asymptotic mean squared error for the estimation of E_c. Let $(Z_1, Y_1), \ldots, (Z_n, Y_n)$, with $Z_i = \Sigma^{-1/2}(X_i - \mathbb{E}[X])$, be an i.i.d. sample from model (1). To measure the estimation error, we define the quantity

$$\mathrm{MSE} = \mathbb{E}\left[\|P_c - \widehat{P}_c\|_F^2\right], \tag{7}$$

where $\|\cdot\|_F$ stands for the Frobenius norm and $\widehat{P}_c$ is derived from the family of vector $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_d)$ defined as

$$\widehat{\eta}_k = \frac{1}{n}\sum_{i=1}^{n} Z_i\psi_k(Y_i), \quad \text{with } \psi_k(Y) = (\mathbb{1}_{\{Y\in I(1)\}}, \ldots, \mathbb{1}_{\{Y\in I(H)\}})\alpha_k = \mathbb{1}_Y^T\alpha_k,$$

and $\alpha_k \in \mathbb{R}^H$. Besides, we introduce $\eta = (\eta_1, \ldots, \eta_d)$ with $\eta_k = \mathbb{E}[Z\psi_k(Y)]$. Consequently, we aim at minimizing MSE according to the family $(\psi_k)_{1\le k\le d}$, or equivalently according to the matrix $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathbb{R}^{H\times d}$. Moreover, since we have

$$\mathrm{MSE} = \mathbb{E}[d - \widehat{d}] + 2\mathbb{E}[\mathrm{tr}(Q_c\widehat{P}_c)], \tag{8}$$

and we suppose that d is known, the minimization of MSE relies only on the minimization of the second term in the previous equality. Hence, this naturally leads to the minimization problem

$$\min_{\alpha} \lim_{n\to+\infty} n\mathbb{E}[\mathrm{tr}(Q_c\widehat{P}_c)],$$

under the constraint of orthogonality of the family $(\eta_k)_{1\le k\le d}$. For the sake of clarity, we prefer to minimize the expectation of the limit in distribution, instead of the limit of the expectation when n goes to infinity, of the sequence $n\,\mathrm{tr}(Q_c\widehat{P}_c)$. To set out clearly the next proposition, let us introduce some notations. Define the matrices $C \in \mathbb{R}^{p\times H}$, $D \in \mathbb{R}^{H\times H}$, such that

$$C = (C_1, \ldots, C_H) \quad \text{with } C_h = \mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}],$$
$$D = \mathrm{diag}_h d_h \quad \text{with } d_h = \left(\mathbb{E}[\|Q_cZ\|^2\mathbb{1}_{\{Y\in I(h)\}}]\right),$$

and

$$G = D^{-\frac{1}{2}}C^TCD^{-\frac{1}{2}}.$$

The matrix G is the Gram matrix of the vector family $(C_h/\sqrt{d_h})_{1\le h\le H}$, Theorem 3 and Remark 1 ensure that its rank is equal to d. Besides, G is diagonalizable and so we define $V = (V_1, V_2) \in \mathbb{R}^{p\times(d+(p-d))}$ such that

$$V^TGV = \begin{pmatrix} D_0 & 0 \\ 0 & 0 \end{pmatrix},$$

where $D_0 \in \mathbb{R}^{d\times d}$.

Proposition 4. *If Z has a finite second order moment, then the random variable $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ has a limit in law W_α as $n \to +\infty$. The minimization problem*

$$\min_{\alpha} \mathbb{E}[W_\alpha] \quad \textit{u.c. } \eta^T\eta = Id,$$

has a unique solution, up to orthogonal transformations, given by

$$\alpha = D^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}.$$

To make a link with other methods and facilitate the programming of OF1, let us express the solution in another way. Instead of expressing the solution in terms of weights α_k's assigned to the indicator functions, we express it in terms of the vectors η_k's associated with these weights. Since the set of functions associated with OF1 is invariant by orthogonal transformations, we choose $\alpha = D^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}$ to simplify the next calculation. We have

$$D^{-\frac{1}{2}}C^TCD^{-\frac{1}{2}}V_1 = V_1D_0,$$

multiplying by $CD^{-\frac{1}{2}}$ on the left and by $D_0^{-\frac{1}{2}}$ on the right, this gives

$$CD^{-1}C^TCD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}} = CD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}D_0.$$

Defining the particular order 1 test function matrix $\widetilde{M}_{\mathrm{OF1}} = CD^{-1}C^T$, and noticing that $\eta = CD^{-\frac{1}{2}}V_1D_0^{-\frac{1}{2}}$, the previous equation is equivalent to

$$\widetilde{M}_{\mathrm{OF1}}\eta = \eta D_0.$$

Thus, since $\widetilde{M}_{\mathrm{OF1}}$ has the same rank as G, we have shown that the vectors η_k's deriving from the optimal weighted family, are the eigenvectors of $\widetilde{M}_{\mathrm{OF1}}$ associated with the nonzero eigenvalues. Besides, it is easy to verify that the previous development is still true when each quantity is replaced by its estimate. Therefore, OF1 relies on the eigendecomposition of an estimator of the matrix $\widetilde{M}_{\mathrm{OF1}}$, whereas SIR is obtained through an eigendecomposition of the matrix $\widetilde{M}_{\mathrm{SIR}}$. To compare both methods, we write their expressions as follows

$$\widetilde{M}_{\mathrm{SIR}} = \sum_{h=1}^{H}\frac{C_hC_h^T}{p_h}, \qquad \widetilde{M}_{\mathrm{OF1}} = \sum_{h=1}^{H}\frac{C_hC_h^T}{d_h}. \tag{9}$$

Hence, SIR and OF1 are closely related because both methods try to obtain the space generated by the C_h's through some PCA. This information seems to be collected more rapidly with OF1 because it minimizes the criterion (7), and as a consequence the convergence rate would be better. This idea is supported by the expression of $\widetilde{M}_{\mathrm{OF1}}$ in which bad slices are less weighted. While $\widetilde{M}_{\mathrm{SIR}} \underset{H\to+\infty}{\to} M_{\mathrm{SIR}}$, $\widetilde{M}_{\mathrm{OF1}}$ converges to

$$M_{\mathrm{OF1}} = \mathbb{E}\left[Z\frac{\mathbb{E}[Z|Y]}{\mathbb{E}[\|Q_cZ\|^2|Y]}\right].$$

As a consequence, OF1 requires the knowledge of Q_c. Therefore we set out a plug-in method to compute Q_c.

OF1 algorithm:

(0) Standardization of X into Z. Initialize $\widehat{Q}_c = I$.
(1) Compute

$$\widehat{d}_h = \frac{1}{n}\sum_{i=1}^{n} \|\widehat{Q}_c Z_i\|^2 \mathbb{1}_{\{Y_i \in I(h)\}}, \qquad \widehat{C}_h = \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{1}_{\{Y_i \in I(h)\}}$$

$$\text{and} \quad \widehat{M} = \sum_{h=1}^{H} \frac{\widehat{C}_h \widehat{C}_h^T}{\widehat{d}_h}.$$

(2) Extract $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_d)$: the d eigenvectors of $\widehat{M}$ with largest eigenvalues.
(3) $\widehat{Q}_c = I - \widehat{\eta}\widehat{\eta}^T$.

Steps 1–3 are repeated until convergence is achieved and then $\widehat{\eta}$ is the estimated basis of the standardized CS derived from OF1. The estimated directions of the CS are $\Sigma^{-\frac{1}{2}}\widehat{\eta}$. At the end of the paper, OF1 is compared to SIR through some simulations.

Naturally the previous development can be carried out with some other total family of functions than indicators, say $\Psi_H = (\psi_1, \ldots, \psi_H)$. The calculation is quite similar, assuming that each ψ_h belong to $L_2(\|Z\|^2)$, the optimization leads to an analogous solution than previously replacing $\widetilde{M}_{\text{OF1}}$ by the matrix $CD_{\Psi_H}C^T$, with $D_{\Psi_H} = \mathbb{E}[\|Q_c Z\|^2 \Psi_H \Psi_H^T]$.

4. Order 2 test function

Basically, TF2 relies on the same approach as TF1 with the difference that it involves higher conditional moments of Z knowing Y. Indeed, we are interested in the space generated by the column vectors of the matrix $\mathbb{E}[ZZ^T\psi(Y)]$ where ψ denotes a measurable function. The following issues are addressed: we first investigate the exhaustiveness of TF2, especially we propose some conditions on ψ that guarantee a comprehensive estimate of the CS, then we look for optimality by introducing OF2.

Let us start with a known fact often presented as the SIR pathology. Consider the regression model

$$Y = g(Z^{(1)}, Z^{(2)}, e), \tag{10}$$

where $e \perp\!\!\!\perp Z \in \mathbb{R}^p$ and g is symmetric with respect to its first coordinate. Assume also that $(Z^{(1)}, Z^{(2)}) \overset{d}{=} (-Z^{(1)}, Z^{(2)})$. Then thanks to the linearity condition we have $Q_c\mathbb{E}[Z\psi(Y)] = 0$ whereas the previous considerations clearly imply that $\mathbb{E}[Z^{(1)}\psi(Y)] = \mathbb{E}[-Z^{(1)}\psi(Y)]$. Therefore for any measurable function ψ, we have that $\mathbb{E}[Z\psi(Y)] = \mathbb{E}[(0, Z^{(2)}, 0, \ldots, 0)^T\psi(Y)]$ and consequently the first direction $(1, 0, \ldots, 0)^T$ cannot be reached by any method based on the inverse regression curve. Clearly, TF1 is sensitive to the SIR pathology. Facing this difficulty, an idea developed first in [16,10] is to explore some higher conditional moments of Z given Y. Thus methods as SIR-II, SAVE, CR, or DR are interested in some properties of the matrix $\mathbb{E}[ZZ^T|Y]$. It is also the case for TF2. Nevertheless we do not follow the same path as other order 2 methods, especially regarding the assumptions required to explore this second order moment. Order 2 methods usually assume that Z has a spherical distribution or at least satisfies the linearity condition, and secondly that $\text{var}(Z|P_cZ)$ is constant, i.e. CCV. In [1, Proposition B.1], stated in Appendix B, shows how strong are the last two assumptions. Accordingly, the assumptions required for order 2 methods are really close to the assumption of normality on the distribution of the predictors. TF2 works under weaker conditions. Actually, the CCV condition is no longer needed and we substitute it with the following one.

Assumption 4 (*Diagonal Conditional Variance (DCV)*)**.**

$$\text{var}(Z|P_cZ) = \lambda_\omega^* Q_c \quad \text{a.s.},$$

with λ_ω^* a real random variable.

To facilitate future proofs and to clear up such a condition we provide an equivalent form in the following lemma.

Lemma 5. *Assume that Z has a finite second moment. Then the following assertions are equivalent,*

(1) *for any orthogonal transformation H such that $HP_c = P_c$, we have*

$$\text{var}(Z|P_cZ) = \text{var}(HZ|P_cZ),$$

(2) *there exists λ_ω^* a real random variable such that* $\text{var}(Z|P_cZ) = \lambda_\omega^* Q_c$.

Moreover, under the linearity condition, we have $\lambda_\omega^* = \frac{1}{p-d}\mathbb{E}\left[\|Q_cZ\|^2|P_cZ\right]$.

Remark 2. Proposition B.1 indicates that coupling CCV and the spherical assumption is equivalent to the normality assumption for Z, which is quite restrictive. In our framework, since sphericity implies DCV, we alleviate this strong link between order 2 methods and the Gaussian assumption. Indeed, if Z is spherical, then its distribution is invariant by any orthogonal transformation, and we have for any measurable function f and for any orthogonal matrix H,

$$\mathbb{E}[ZZ^T f(P_c Z)] = \mathbb{E}[HZZ^T H^T f(P_c HZ)].$$

In particular, the previous equation is true for any H which leaves invariant vectors of E_c and we obtain (1) of Lemma 5 which is equivalent to DCV. Thus, we have just proved that the spherical assumption implies DCV.

The following theorem is the analogous of Theorem 2 for TF2. We define

$$M_\psi = \mathbb{E}[ZZ^T \psi(Y)] \quad \text{and} \quad \lambda^*_\psi = \frac{1}{p-d}\mathbb{E}\left[\|Q_c Z\|^2 \psi(Y)\right].$$

Theorem 6. *Assume that Z satisfies Assumptions 1 and 4 and has a finite second moment. Then, for every measurable function $\psi \in L_1(\|Z\|^2)$, we have*

$$\operatorname{span}(M_\psi - \lambda^*_\psi I) \subset E_c.$$

In practice, because λ^*_ψ is unknown, it seems difficult to use Theorem 6. Nevertheless, we do not need to know this particular eigenvalue, this issue is addressed in Remark 3. Besides, a consequence of Theorem 6 is that $E_c^\perp$ is included in the eigenspace of the matrix M_ψ associated with the eigenvalue λ^*_ψ. Therefore, if all the other eigenvalues are different from λ^*_ψ, the eigenspace associated with λ^*_ψ is equal to $E_c^\perp$. If this is true, the inclusion in Theorem 6 becomes an equality, i.e. all the directions of E_c could be recovered. This idea has a central place in the next section where this eigenvalue problem is addressed.

4.1. Exhaustiveness for TF2

An important tool in this section is the eigendecomposition of the matrix M_ψ, therefore we try to be more clear in introducing the following notations. Let λ_ψ and λ_Y be two functions $\mathbb{R}^p \to \mathbb{R}$ respectively defined by

$$\lambda_\psi(\eta) = \mathbb{E}[(\eta^T Z)^2 \psi(Y)] \quad \text{and} \quad \lambda_Y(\eta) = \mathbb{E}[(\eta^T Z)^2 | Y],$$

for every $\eta \in \mathbb{R}^p$. Notice that if η is a unit eigenvector of M_ψ (resp. $\mathbb{E}[ZZ^T|Y]$), then $\lambda_\psi(\eta)$ (resp. $\lambda_Y(\eta)$) is equal to the eigenvalue of the matrix M_ψ (resp. $\mathbb{E}[ZZ^T|Y]$) associated with η. However, recalling that $E_c^\perp$ is included in an eigenspace of M_ψ and $\mathbb{E}[ZZ^T|Y]$, the functions λ_ψ and λ_Y are both constant on the centered spheres of $E_c^\perp$. Their respective values on the unit sphere of $E_c^\perp$ are noted λ^*_ψ and λ^*_Y.

Definition. Let ψ be a measurable function, we call ψ-space the vector space of $\mathbb{R}^p$

$$E_\psi = \operatorname{span}(M_\psi - \lambda^*_\psi) = \operatorname{span}\left(\eta \in B(0,1) \subset \mathbb{R}^p, M_\psi \eta = \lambda^*_\psi \eta\right)^\perp.$$

Theorem 6 indicates that any ψ-space is included in E_c. However, there is no guarantee of the existence of a ψ-space equal to E_c. We follow the same path as for TF1, i.e. we consider some transformations of Y belonging to a dense family. Nevertheless, the results are quite different because we provide the existence of a single function ψ such that $E_\psi = E_c$. A unique additional assumption is needed.

Assumption 5 (*Order 2 Coverage Condition*)**.**

$$\forall \eta \in E_c, \quad \|\eta\| = 1 \qquad \mathbb{P}\left(\mathbb{E}\left[(\eta^T Z)^2 | Y\right] = \mathbb{E}\left[\left.\frac{\|Q_c Z\|^2}{p-d}\right| Y\right]\right) < 1.$$

Assumption 5 reflects some similarities with other work such as [20,19]. As highlighted in Remark 2, our set of assumptions is weaker than their because DCV has replaced CCV. To match their context, assume that CCV is satisfied. Then, Assumption 5 becomes "$\mathbb{E}[(\eta^T Z)^2|Y]$ is nondegenerate", i.e. is not a constant almost surely. Otherwise, TF1 allows an exhaustive estimation of the CS provided that $\mathbb{E}[(\eta^T Z)|Y]$ is nondegenerate. Thus the exhaustiveness condition of TF is the union of the two previous, i.e.

$$\mathbb{E}[(\eta^T Z)^2|Y] \quad \text{or} \quad \mathbb{E}[(\eta^T Z)|Y] \text{ is nondegenerate},$$

which is the same than the one proposed for DR in [19]. Accordingly, TF evolves in a more general context given by DCV but the assumptions ensuring its exhaustiveness are similar. These assumptions can be understood as theoretical ones because they are difficult to check in practice.

Theorem 7. *Assume that Z and Y satisfy Assumptions 1, 4 and 5. Assume also that Z has a finite second moment, then if Ψ is a total countable family in the space $L_1(\|Z\|^2)$, there exists ψ a finite linear combination of functions in Ψ such that*

$$E_\psi = E_c.$$

Theorem 7 highlights some relevant facts about TF2. In addition to providing the existence of a ψ-space equal to E_c, it gives some information about the function ψ to be used. Indeed, Theorem B.2 indicates that the relevant families of functions for TF2 are those that separate the points. Hence, as for TF1, this suggests the use of TF2 with any of these families. For each such family, there exists a function ψ such that $E_\psi = E_c$, yet it does not provide an explicit form of such a ψ. Hence, we set out the following corollary which is the counterpart of Theorem 3 for TF2.

Corollary 8. *Assume that Z and Y satisfy Assumptions 1, 4 and 5. Assume also that Z has a finite second moment then, if Ψ is a total countable family in the space $L_1(\|Z\|^2)$, we have*

$$\bigoplus_{\Psi_H} E_\psi = E_c,$$

where Ψ_H is a finite subset of Ψ.

4.2. Optimality for TF2: OF2

For TF1 we needed at least d functions to recover the CS entirely. For this reason, it was convenient to develop a framework with weighted indicators because it led to a matrix optimization problem. In other words we fixed the class of functions for TF1 to solve a finite dimensional optimization problem. Actually, for TF2 we follow a different path: we choose to optimize over all the measurable functions thanks to Gâteaux derivatives.

We have already highlighted that the eigenvectors of the matrix M_ψ can be decomposed into two blocks: the ones associated with the eigenvalue λ^*_ψ and the others which necessarily belong to E_c. Theorem 7 goes further by arguing that for some ψ, the eigenvectors associated with different eigenvalues than λ^*_ψ generate E_c. Therefore, P_c can be derived from this set of eigenvectors. A natural way to proceed is to estimate each quantity by its empirical version. Recall that $(Z_1, Y_1), \ldots, (Z_n, Y_n)$, with $Z_i = \Sigma^{-1/2}(X_i - \mathbb{E}[X])$, is an i.i.d. sample from model (1). We define

$$\widehat{M}_\psi = \frac{1}{n}\sum_{i=1}^n Z_i Z_i^T \psi(Y_i)$$

and the function $\widehat{\lambda}_\psi : \mathbb{R}^p \to \mathbb{R}$ as $\widehat{\lambda}_\psi(\eta) = \eta^T \widehat{M}_\psi \eta$ for every $\eta \in \mathbb{R}^p$. Since d is assumed to be known, we define the projector $\widehat{P}_c = \widehat{\eta}_\psi \widehat{\eta}_\psi^T$ where $\widehat{\eta}_\psi \in \mathbb{R}^{p\times d}$ are the d eigenvectors of $\widehat{M}_\psi$ associated with the eigenvalues the farthest from λ^*_ψ. Because of the symmetry of the matrix $\widehat{M}_\psi$ and M_ψ, the convergence $\widehat{M}_\psi \xrightarrow{\mathbb{P}} M_\psi$ implies the convergence in probability of the associated eigenvalues (see [13] for some details). As a consequence, one can express the projectors with the Riesz formula. Let $\mathcal{C}$ be a contour of the complex plan which encloses the eigenvalues different from λ^*_ψ. We prefer to work with P_c and its estimator $\widehat{P}_c$ expressed as

$$P_c = \oint_{\mathcal{C}} (Iz - M_\psi)^{-1} dz \quad \text{and} \quad \widehat{P}_c = \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1} dz.$$

Because of Eq. (8), we minimize MSE through the quantity $\mathbb{E}[\mathrm{tr}(Q_c\widehat{P}_c)]$. As we did for OF1, we first calculate the limit in law of the random variable $n\,\mathrm{tr}(Q_c\widehat{P}_c)$, as n goes to infinity and then we derive its expectation. The next proposition is dedicated to this calculation.

Proposition 9. *Let $\psi \in L_2(\|Z\|^4)$ such that $E_\psi = E_c$. Then $n\,\mathrm{tr}(Q_c\widehat{P}_c)$ has a limit in law W_ψ and*

$$\mathbb{E}[W_\psi] = \mathrm{tr}\left(\mathbb{E}\left[ZZ^T\|Q_cZ\|^2\psi(Y)^2\right]P_c(P_cM_\psi - I\lambda^*_\psi)^{-2}\right).$$

The above proposition provides the expression of the quantity to minimize with respect to the function ψ. The next lines are attached to find a minimizer of $\mathbb{E}[W_\psi]$. This informal calculation leads to a fixed point equation whose solution is expected to be a global minimum of $\mathbb{E}[W_\psi]$. Thanks to Proposition 9 the quantity to minimize can be written as

$$\mathbb{E}[W_\psi] = \mathrm{tr}(\mathbb{E}[ZZ^TP_c\|Q_cZ\|^2\psi(Y)^2](P_cM_\psi - I\lambda^*_\psi)^{-2}),$$

or introducing the notations $A = ZZ^TP_c\|Q_cZ\|^2$ and $B = P_cZZ^T - \frac{\|Q_cZ\|^2}{p-d}I$,

$$\mathbb{E}[W_\psi] = \mathrm{tr}\left(\mathbb{E}[A\psi(Y)^2]\,\mathbb{E}[B\psi(Y)]^{-2}\right).$$

Thus we are looking for ψ such that

$$\frac{\partial}{\partial t}\mathbb{E}[W_{\psi+t\delta}]\Big|_{t=0} = 0,$$

for every bounded measurable function δ, or equivalently,

$$\mathbb{E}\left[2\,\mathrm{tr}\left(A\delta\psi\mathbb{E}[B\psi]^{-2}\right) - \mathrm{tr}\left(\mathbb{E}[A\psi^2]\mathbb{E}[B\psi]^{-1}\{B\delta\mathbb{E}[B\psi]^{-1} + \mathbb{E}[B\psi]^{-1}B\delta\}\mathbb{E}[B\psi]^{-1}\right)\right] = 0,$$

where δ and ψ stand for $\delta(Y)$ and $\psi(Y)$. Define the functions $A(Y) = \mathbb{E}[A|Y]$ and $B(Y) = \mathbb{E}[B|Y]$. Since the previous equation is true for any Y-measurable random variable $\delta(Y)$, we obtain

$$2\,\mathrm{tr}\left(A(Y)\psi(Y)\mathbb{E}[B\psi]^{-2}\right) - \mathrm{tr}\left(\mathbb{E}[A\psi^2]\mathbb{E}[B\psi]^{-1}\{B(Y)\mathbb{E}[B\psi]^{-1} + \mathbb{E}[B\psi]^{-1}B(Y)\}\mathbb{E}[B\psi]^{-1}\right) = 0 \quad \text{a.s.},$$

which leads to the implicit equation

$$\psi(y) = \frac{\mathrm{tr}\left(\mathbb{E}[B\psi]^{-1}\mathbb{E}[A\psi^2]\mathbb{E}[B\psi]^{-1}\{\mathbb{E}[B\psi]^{-1}B(y) + B(y)\mathbb{E}[B\psi]^{-1}\}\right)}{2\,\mathrm{tr}(A(y)\mathbb{E}[B\psi]^{-2})}.$$

Since $P_c = \eta_\psi \eta_\psi^T$, we have

$$\mathbb{E}[B\psi]^{-1}\eta_\psi = \eta_\psi D_\psi,$$

where $D_\psi = \mathrm{diag}_k(\lambda_\psi(\eta_k) - \lambda_\psi^*)^{-1}$ and η_k is the k-th column vector of η_ψ. Besides, a simple use of the linearity condition provides that $\mathbb{E}[\eta^T ZZ^T|Y] = \mathbb{E}[\eta^T ZZ^T P_c|Y]$ for every $\eta \in E_c$. Consequently, we have

$$\eta_\psi^T B(y) = \eta_\psi^T B(y) P_c$$

and then, we obtain

$$\psi(y) = \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi\{D_\psi \widetilde{B}(y) + \widetilde{B}(y)D_\psi\}\right)}{2\,\mathrm{tr}\left(\widetilde{A}(y)D_\psi^2\right)},$$

where

$$A_\psi = \mathbb{E}\left[\eta_\psi^T ZZ^T \eta_\psi \|Q_c Z\|^2 \psi(Y)^2\right], \qquad \widetilde{A}(y) = \eta_\psi^T A(y)\eta_\psi, \qquad \widetilde{B}(y) = \eta_\psi^T B(y)\eta_\psi,$$

are $d \times d$ matrices. Using the symmetry of the matrices A_ψ and $\widetilde{B}(y)$, and some well-known properties of the trace, we obtain

$$\psi(y) = \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi \widetilde{B}(y) D_\psi\right)}{\mathrm{tr}\left(\widetilde{A}(y)D_\psi^2\right)}. \tag{11}$$

A solution of Eq. (11) is noted ψ_{OF2}, it is an optimal function inside the TF2 framework with respect to criterion (7). Hence, we define the OF2 matrix as

$$M_{\mathrm{OF2}} = \mathbb{E}[ZZ^T \psi_{\mathrm{OF2}}(Y)].$$

To calculate ψ_{OF2}, we propose an iteration of the fixed point Eq. (11). Before we state a more accurate algorithm to compute OF2, in particular to estimate the matrix M_{OF2}, we need to approximate ψ_{OF2}. Indeed, since $\widetilde{A}$ and $\widetilde{B}$ are unknown functions, one can use a slicing approximation and define $\widetilde{\psi}_{\mathrm{OF2}}$ as a solution of

$$\psi(y) = \sum_h \frac{\mathrm{tr}\left(D_\psi A_\psi D_\psi \widetilde{B}_h D_\psi\right)}{\mathrm{tr}\left(\widetilde{A}_h D_\psi^2\right)} \mathbb{1}_{\{y \in I(h)\}},$$

where $\widetilde{A}_h = \mathbb{E}[\widetilde{A}(Y)\mathbb{1}_{\{y\in I(h)\}}]$ and $\widetilde{B}_h = \mathbb{E}[\widetilde{B}(Y)\mathbb{1}_{\{y\in I(h)\}}]$. Now we set out the OF2 method based on the family of indicator functions. The following algorithm describes the iterations needed to implement our method. For a better understanding, we based the algorithm on the weights α_h's instead of the function $\psi(y) = \sum_h \alpha_h \mathbb{1}_{\{y\in I(h)\}}$. Besides $\widehat{A}_{\widehat{\psi}}$ and $\widehat{D}_{\widehat{\psi}}$ are noted $\widehat{A}$ and $\widehat{D}$, and we will need

$$M_h = \mathbb{E}[ZZ^T \mathbb{1}_{\{Y\in I(h)\}}] \quad \text{and} \quad \lambda_h = \mathbb{E}\left[\frac{\|Q_c Z\|^2}{p-d}\mathbb{1}_{\{Y\in I(h)\}}\right].$$

OF2 algorithm:

(0) Standardization of X into Z. Compute

$$\widehat{M}_h = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T \mathbb{1}_{\{Y_i \in I(h)\}}, \qquad \widehat{\lambda}_h = \text{median}(\lambda \in \text{spectrum}(\widehat{M}_h)),$$

and initialize $\widehat{\alpha}_h \overset{d}{=} \mathcal{U}[0,1]$ for every $h = 1, \dots, H$.

(1) Identify[1] the eigenvectors $\widehat{\eta} = (\widehat{\eta}_1, \dots, \widehat{\eta}_d) \in E_c$ of $\widehat{M} = \sum_h \widehat{\alpha}_h \widehat{M}_h$.

(2) Derive $\widehat{D} = \text{diag}_k(\widehat{\lambda}_{\widehat{\psi}}(\widehat{\eta}_k) - \widehat{\lambda}^*_{\widehat{\psi}})^{-1}$ with $\widehat{\psi}(y) = \sum_h \widehat{\alpha}_h \mathbb{1}_{\{y \in I(h)\}}$, $\widehat{Q}_c = I - \widehat{\eta}\widehat{\eta}^T$ and

$$\widehat{A} = \sum_h \widehat{\alpha}_h \widehat{\eta}^T \widehat{A}_h \widehat{\eta}, \quad \text{with } \widehat{A}_h = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i^T \|\widehat{Q}_c Z_i\|^2 \mathbb{1}_{\{Y_i \in I(h)\}}.$$

(3) Compute

$$\widehat{\alpha}_h = \frac{\text{tr}\left(\widehat{D}^2 \widehat{A} \widehat{D}\ (\widehat{\eta}^T \widehat{M}_h \widehat{\eta} - \widehat{\lambda}_h I)\right)}{\text{tr}\left(\widehat{\eta}^T \widehat{A}_h \widehat{\eta}\, \widehat{D}^2\right)}.$$

Repeat the last three steps until the convergence is achieved. The resulting function $\widehat{\psi}_{\text{OF2}}$ is an estimate of the function ψ_{OF2}. Finally the set of vectors $\widehat{\eta}$ forms an estimated basis of the standardized CS. The space generated by $\Sigma^{-\frac{1}{2}}\widehat{\eta}$ provides an estimate of E_c by OF2.

Remark 3. An important practical issue for TF2 and in particular for OF2 is the way we identify the eigenvectors of $\widehat{M}_\psi$ that converge to some vectors of E_c or equivalently the way we identify their associated eigenvalues. This intervenes at each iteration of our algorithm to estimate D_ψ and η_ψ. Although λ^*_ψ is unknown, the theoretical background of TF2 advocates for an identification process based on the eigenvalues. Indeed, as it is pointed out by Theorem 6, the eigenvalues of M_ψ associated with eigenvectors included in $E_c^\perp$ are equal. We built an algorithm based on this fact but it was not sufficiently robust to small samples. We thus prefer to develop another one which takes into account the eigenvectors of $\widehat{M}_\psi$. Let η be an eigenvector of $\widehat{M}_\psi$, the identification process is based on a measure of the dependence between $\eta^T Z$ and Y. More precisely, we consider the Pearson's chi-squared statistic of the test of independence between $\eta^T Z$ and Y. Therefore, for each eigenvector we divide the range of $\eta^T Z$ into H slices noted $J(h)$ and we calculate

$$S(\eta) = \sum_{h,h'} \frac{\left(p_{hh'} - \overline{p_{hh'}}^h \overline{p_{hh'}}^{h'}\right)^2}{\overline{p_{hh'}}^h \overline{p_{hh'}}^{h'}}$$

where $p_{h,h'} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{Y_i \in I(h)\}} \mathbb{1}_{\{(\eta^T Z_i) \in J(h')\}}$ and $\overline{\cdot}^h$ is the mean over h. Then the d eigenvectors of $\widehat{M}_\psi$ having the highest values of S are identified as converging in E_c. As a consequence, at step 2 of the OF2 algorithm, the $\widehat{\lambda}_{\widehat{\psi}}(\widehat{\eta}_k)$'s are the eigenvalues of $\widehat{M}$ associated with the eigenvectors $\widehat{\eta}_k$'s with the d highest values of S, $\lambda^*_{\widehat{\psi}}$ is the median over the other eigenvalues. In Section 6, we performed OF2 with this algorithm.

5. Estimation of the dimension

All along the article, the dimension of the CS was assumed to be known. Its estimation is a crucial point in SDR since it corresponds to the number of explicative variables we keep in the regression. Clearly if the dimension is underestimated, then we loose some information about the response, and on the contrary we cannot get the suitable nonparametric convergence rates for the estimation of the regression function. We raise this issue for TF1 and TF2. The estimation of d can be reasonably conducted after the estimation of the matrix of interest, say M, in the following way. As we pointed out before, under some conditions, one can get

$$\text{span}(M) = E_c,$$

and clearly, the estimation of d amounts to estimate the rank of M. Actually, to estimate the rank of such matrix, one can use the hypothesis testing methodology proposed by Li [16] whose null hypothesis is

$$H_0 : d = m \quad \text{against } H_1 : d > m,$$

[1] See Remark 3 for some details about this point.

where d stands for the true dimension. Then we start by testing $d = 0$ against $d > 0$ which can be seen as a test for the existence of a DRS. If it is rejected we go a step further $m := m + 1$ until the first acceptance. If $d = m$ is accepted, then m is an estimate of the dimension of E_c. The usual statistic employed in SDR is

$$\widehat{\Lambda} = n \sum_{k=1}^{p-m} \widehat{\lambda}_k^2$$

where $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)$ are the singular values of an estimator of M arranged in ascending order. Roughly speaking, the statistic goes to infinity under H_1 because at least one of the eigenvalues goes to a positive constant. Under H_0 and some mild conditions $\widehat{\Lambda}$ converges in law. This is the issue raised by Theorem 1 in [4], stated in Appendix B as Theorem B.3. Thanks to this theorem, most of the SDR methods can provide an estimate of the dimension of E_c. For SIR, because $\widetilde{M}_{\mathrm{SIR}} = C\,\mathrm{diag}_h(p_h^{-1})C^T$, it is preferable to apply Theorem B.3 directly with the matrix $\mathrm{diag}_h(p_h)^{-1/2}C$, then we define $\widehat{\Lambda}_{\mathrm{SIR}}$ as $\widehat{\Lambda}$ with $M = C\,\mathrm{diag}_h(p_h^{-1/2})$. Because of the unknown asymptotic distribution of $\widehat{\Lambda}$ under H_0 in general, it is interesting to study the behavior of the statistic $\widehat{\Lambda}$ under some usual SDR assumptions in order to take advantage of the substantial simplifications they involve. For instance, [3] show that under the linearity condition and CCV, $\widehat{\Lambda}_{\mathrm{SIR}}$ is asymptotically chi-squared. Hence in the following, we provide the asymptotic distribution of $\widehat{\Lambda}$ in a general TF1 context without specifying the family of function $\Psi_H = (\psi_1, \ldots, \psi_H)^T$. Moreover our study involves both sets of assumptions: CCV and DCV (see Remark 2 for details about such assumptions). We use a parametrization quite similar to that of Section 3.2 by defining the matrix

$$M_\alpha = C\alpha(C\alpha)^T,$$

with $\alpha \in \mathbb{R}^{H\times H}$ could be unknown, $C = (C_1, \ldots, C_H)$, and $C_h = \mathbb{E}[Z\psi_h(Y)]$. Define also U_0 and V_0 as the respective basis of the left and right singular spaces of the matrix $C\alpha$ associated with the singular value 0. Assume that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is an i.i.d. sample from model (1) and define

$$\widehat{Z}_i = \widehat{\Sigma}^{-1/2}(X_i - \overline{X}),$$

with $\overline{\cdot}$ the empirical mean. Then we can define the estimator

$$\widehat{M}_\alpha = \widehat{C}\widehat{\alpha}(\widehat{C}\widehat{\alpha})^T,$$

where $\widehat{C} = (\widehat{C}_1, \ldots, \widehat{C}_H)$, $\widehat{C}_h = \frac{1}{n}\sum_{i=1}^n \widehat{Z}_i\psi_h(Y_i)$, and $\widehat{\alpha} \in \mathbb{R}^{H\times H}$ is an estimator of α. The next theorem studies the asymptotic distribution of

$$\widehat{\Lambda}_{\mathrm{TF1}} = n \sum_{k=1}^{p-m} \widehat{\lambda}_k^2,$$

where $(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_p)$ are the singular values of $\widehat{C}\widehat{\alpha}$ arranged in ascending order.

Theorem 10. *Under H_0, assume that Z satisfies Assumptions 1 and 4 (resp. 1 and 2) and has a finite second moment, then if $\psi_h \in L_2(\|Z\|^2)$ and $\sqrt{n}(\widehat{C}\widehat{\alpha} - C\alpha)$ has an asymptotic Gaussian distribution, we have*

$$\widehat{\Lambda}_{TF1} \xrightarrow{d} \sum_{k=1}^{H-d} \omega_k \xi_k,$$

where the ξ_k's are i.i.d. chi-squared variables with $p - d$ degrees of freedom and the ω_k's are the eigenvalues of the matrix $V_0^T\alpha^T\Delta\alpha V_0$ where

$$\Delta = \mathbb{E}\left[(p-d)^{-1}\|Q_c Z\|^2(\Psi_H(Y) - \mathbb{E}[\Psi_H(Y)])(\Psi_H(Y) - \mathbb{E}[\Psi_H(Y)])^T\right] \quad (\textit{resp. } \Delta = \mathrm{var}(\Psi_H(Y))).$$

The above theorem is a general statement about the estimation of the dimension of E_c for TF1. Notice that the framework employed contains SIR and OF1 as special cases. We highlight in the following some relevant applications. Under CCV, considering the indicator functions and taking $\alpha = \mathrm{diag}_h p_h^{-1}$, we obtain the same result as [3, Corollary 1], regarding M_{SIR}. Besides, it is easy to show that CCV implies that $d_h = p_h(p - d)$, then if $\alpha = \mathrm{diag}_h d_h^{-1}$, we provide the asymptotic law of $\widehat{\Lambda}_{\mathrm{TF1}}$ for OF1, i.e.

$$\widehat{\Lambda}_{\mathrm{OF1}} \xrightarrow{d} (p-d)^{-1}\chi^2_{(p-d)(H-d-1)}.$$

The above convergence highlights that, as SIR, OF1 provides a pivotal test for the considered statistic under CCV.

In general the asymptotic distribution of $\widehat{\Lambda}_{\mathrm{TF1}}$ is no longer chi-squared and the weights ω_k's need to be estimated. Theorem 10 emphasizes a pivotal version of such a test for any family of functions thanks to a good specification of the matrix α. For clarity assume that Δ is a full rank matrix, one can take $\alpha = \Delta^{-1/2}$ in Theorem 10 under DCV or CCV. We get for both

$$\widehat{\Lambda}_{\mathrm{TF1}} \xrightarrow{d} \chi^2_{(p-d)(H-d)},$$

where α can be respectively estimated by

$$\frac{1}{n(p-d)}\sum_{i=1}^{n}\|\widehat{Q}_c\widehat{Z}_i\|^2(\Psi_H(Y_i)-\overline{\Psi_H})(\Psi_H(Y_i)-\overline{\Psi_H})^T \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}(\Psi_H(Y_i)-\overline{\Psi_H})(\Psi_H(Y_i)-\overline{\Psi_H})^T,$$

where $\widehat{Q}_c$ is estimated from the considered TF1 method. Taking advantage of the SDR context, this kind of approach goes in the sense of the Wald-type pivotal statistic studied for instance in [4].

Using the same approach, it is possible to obtain the asymptotic distribution of such a statistic for TF2. Nevertheless, such matrices are not positive and then the test needs to be based on the sum of squares of the eigenvalues of M_{TF2}. In this case, the eigenvalues ω_k's in Theorem B.3 are more complicated than for TF1 even if we assume DCV or CCV. As a consequence it seems less attractive to follow the same path as previously. However one could follow [4, Theorem 1], to provide a consistent test, assuming sufficient finite moments for Z in order to ensure the convergence of $\widehat{\Lambda}$ on the one hand, and in order to estimate consistently the weights ω_k's on the other hand.

6. Simulations

In this section, we evaluate OF1, OF2 and some other SDR methods through different regression models. We first compare OF1 with SIR and IRE and then, we compare OF2 to some order 2 methods through pathological models for order 1 methods (see Example 10). To measure the performance of a method we evaluate the estimation error with the following distance: for two subspace E_1 and E_2, if P_1 and P_2 are their respective orthogonal projectors, the distance between E_1 and E_2 is

$$\text{Dist}(E_1, E_2) = \|P_1 - P_2\|_F. \tag{12}$$

In the following study, each method is evaluated for a single model. Each boxplot is based on 100 runs of the considered model. All along the simulation study, in order to appreciate the real intrinsic quality of each method, we assume that the variance and the mean of the predictors are known. As a consequence we do not take into account the bias introduced by poor estimates of the variance and the mean. Besides, we compare the distance (12) between the estimated standardized directions and the standardized CS.

For each method, when the response is continuous, we discretize its range into H slices, each containing the same number of observations. Both methods OF1 and OF2 require the iteration of the so called OF1 and OF2 algorithms (see Sections 3.2 and 4.2). In each case, the number of iterations equals 5. Finally, this simulation study is organized according to four examples that combine different distributions for the predictors.

6.1. OF1 and order 1 methods

The order 1 methods we computed include SIR and IRE. Let us consider the case where the predictors have a Gaussian distribution. Clearly P_cZ and Q_cZ are two independent random vectors and then $\mathbb{E}[\|Q_cZ\|^2|Y] = \mathbb{E}[\mathbb{E}[\|Q_cZ\|^2|P_cZ]|Y] = p-d$. Therefore $\text{span}(M_{\text{OF1}}) = \text{span}(M_{\text{SIR}})$ and OF1 is similar to SIR. Simulations made in this case highlight the similarity between the selected methods and are not presented here. Besides, to reach a point of view developed in the simulation study of [9], we are interested in the link between the variations of $\text{var}(Z|Y)$ and the performance of the presented methods. Clearly, according to Eq. (9), the variations of the random variable $\mathbb{E}[\|Q_cZ\|^2|Y]$ emphasize the differences between SIR and OF1. Indeed if this one is a constant, then $d_h = \mathbb{E}[\|Q_cZ\|^2\mathbb{1}_{\{Y\in I(h)\}}] = (p-d)p_h$ and OF1 is the same method as SIR. Consequently, SIR estimates are near optimal with respect to criterion (7) when the variations of $\mathbb{E}[\|Q_cZ\|^2|Y]$ are near 0. Besides, if this random variable is nonconstant then also the d_h's and the differences between both methods are emphasized. Moreover, the random variables $\mathbb{E}[\|Q_cZ\|^2|Y]$ and $\text{var}(Z|Y)$ are strongly linked, and as it was the case to distinguish IRE from SIR, the variations of $\text{var}(Z|Y)$ play an important role to differentiate OF1 from SIR. Consequently, to point out the differences between these methods, we generate non-Gaussian predictors in the following two examples.

Example 1. Let $N_1 \in \mathbb{R}^p, N_2 \in \mathbb{R}^p$ be two independent standard Gaussian vectors, let ϵ be a Bernoulli random variable with mean 1/2. The predictor vector $X = (X^{(1)}, \ldots, X^{(p)})$ is generated as a Gaussian mixture through the equation

$$X = (\mu_1 + \sigma_1 N_1)\epsilon + N_2(1-\epsilon),$$

and it would be interesting to consider different values of $\sigma_1 \in \mathbb{R}$ and $\mu_1 \in \mathbb{R}^p$. We introduce the following models

$$\text{Model I: } Y = \tanh(X^{(1)}/3) + 0.1e$$
$$\text{Model II: } Y = X^{(2)}|1 + X^{(1)}/3| + e$$

where $e \overset{d}{=} \mathcal{N}(0,1)$. For Model I, an interesting parametrization is

$$\mu_1 = (a, 0, \ldots, 0)^T,$$

and then we can consider different values for a and σ_1. Such a distribution for the predictors induces two regimes. To highlight differences between both regimes respectively determined by $\epsilon = 1$ and $\epsilon = 0$, one can take the parameter a

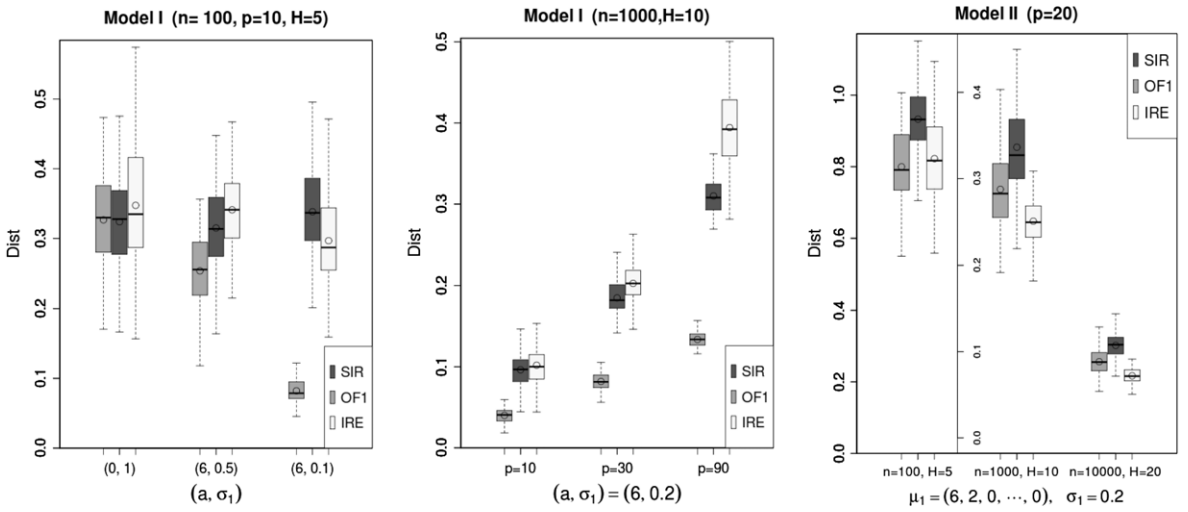

Fig. 1. Plot of the distance error for OF1, SIR and IRE in Example 1.

far from 0 and $\sigma_1 \neq 1$, say $a = 6$ and $\sigma_1 = 0.5$. Clearly the C_h's corresponding to small Y, have more chances to come from the second regime $\epsilon = 0$, which induces a poor estimate for such C_h's. On the contrary, the other C_h's tend to be well estimated. In this case the error of the SIR method is to uniformly weight these slices whereas OF1 does not. To be more comprehensive, we compute the methods with different parametrizations. The boxplots and the averages of the distance (12) between the standardized CS and its estimates over 100 simulated samples are given in Fig. 1. With the same model, in this figure, we also provide a graph to describe the effect of an increase of p.

For Model II, E_c has dimension 2 and then is more difficult to estimate. We consider

$$\mu_1 = (6, 2, \ldots, 0)^T,$$

and essentially, Model II provides similar graphs and interpretations as Model I. As a result, we analyze through this model the impact of an increase of n. The corresponding graph has been included in Fig. 1.

For each model and in all the parameter configurations, OF1 performs better than SIR. Between OF1 and IRE, the conclusion is quite a lot more mitigated. The chosen configurations reflect different kinds of difficulties. The situation presented in the first graph reflects a too small sample number $n = 100$ with respect to $p = 10$ to provide a good estimate. When $(\mu_1, \sigma_1) = (0, 1)$, the predictors are normally distributed and there are no significant differences between the methods. By increasing μ_1 and reducing σ_1, we move away from the Gaussian assumption and OF1 is the only one to improve its accuracy. Indeed, OF1 performs better than SIR and IRE around 86% of the time when $(\mu_1, \sigma_1) = (6, 0.5)$ and 100% of the time when $(\mu_1, \sigma_1) = (6, 0.1)$. Besides, the second graph shows that OF1 is more robust to a high dimensional set-up. The most sensitive method to the increase of p is IRE because it requires the estimation of a large matrix. Finally, the last graph emphasizes that IRE is the most accurate when n is large.

Example 2. This example is interesting because it includes logistic models in the SDR framework. It is inspired from [9], Model A. We generalize their model by introducing some noise as described in the following. Let ϵ be a real random variable uniformly distributed on $\{1, 2, 3\}$, and let $N_1 \in \mathbb{R}^p$, $N_2 \in \mathbb{R}^p$, $N_3 \in \mathbb{R}^p$ be independent Gaussian vectors with respective moments $(\mu_1\mathbf{1}, \sigma_1^2 I)$, $(\mu_2\mathbf{1}, \sigma_2^2 I)$ and $(\mu_3\mathbf{1}, \sigma_3^2 I)$ where $\mathbf{1} = (1, \ldots, 1)^T$. The vector X is generated as a Gaussian mixture through the equation

$$X = N_1\mathbb{1}_{\{\epsilon=1\}} + N_2\mathbb{1}_{\{\epsilon=2\}} + N_3\mathbb{1}_{\{\epsilon=3\}},$$

and Y with the proportional-odds model defined by

$$\text{Model III: } Y = \sum_{j=1}^{3} j\mathbb{1}_{\{\pi_{j-1} \leq U \leq \pi_j\}},$$

with $U \overset{d}{=} \mathcal{U}([0, 1])$ and the cumulative probability functions

$$\pi_0 = 0, \qquad \pi_1 = \frac{\exp(\theta_1 - \mathbf{1}^T X)}{1 + \exp(\theta_1 - \mathbf{1}^T X)}, \qquad \pi_2 = \frac{\exp(\theta_2 - \mathbf{1}^T X)}{1 + \exp(\theta_2 - \mathbf{1}^T X)}, \qquad \pi_3 = 1.$$

First note that Model III implies that $Y = f(\mathbf{1}^T X, U)$ and as a consequence the CS exists for this kind of models. In our case, the CS is generated by the vector $\mathbf{1}$ and the CS is equal to the standardized CS. For clarity, we prefer to work with the mean

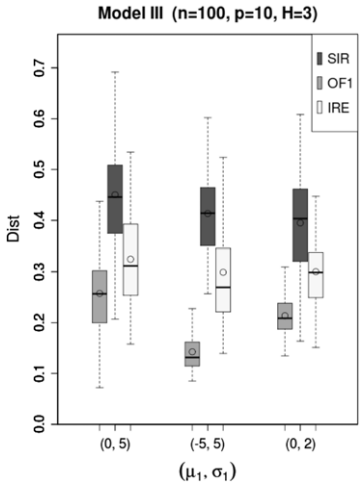

Fig. 2. Plot of the distance error for OF1 and SIR in Example 2.

and the standard error of the predictors divided respectively by p and $\sqrt{p}$ so that the mean and the standard error of $\mathbf{1}^T X$ do not depend on p. Working with the new scaled parameters, we fix $\mu_2 = 5$, $\mu_3 = 8$, $\sigma_2 = 0.5$, and $\sigma_3 = 0.5$. Then we can specify the cumulative probability functions by taking $\theta_1 = 3.5$ and $\theta_2 = 6.5$, so that it is realistic with respect to the means μ_2 and μ_3. To visualize such a model, one could draw in the same plot the cumulative probability functions π_1, $\pi_2 - \pi_1$, and $1 - \pi_2$, and the density of $\mathbf{1}^T X$. Each state of the response tends to correspond to some regime of the Gaussian mixture. The parameter H is fixed to 3, the number of states of the response. In Fig. 2, we test the accuracy of OF1, SIR and IRE facing Model III for different configurations of the parameters μ_1 and σ_1. The dimension p and the sample number n have been taken to provide neither a simple situation, nor a too difficult one.

In Fig. 2, the presented graph starts by a model with a lot of noise. The second and third situation reflects respectively a shift of the mean μ_1 and a shift of the variance σ_1. In each case, this reduces the noise and the estimation of the CS is more accurate for all the methods. Again when some estimated $\widehat{C}_h$'s have a small variance, OF1 manages to take advantage of the situation.

6.2. OF2 and order 2 methods

We compare several well-known order 2 dimension reduction methods with OF2. Order 2 methods we computed include SAVE, pHd, SIR-II and DR. For the considered models, pHd does not work as well as the others. Therefore we focus on a comparison between SAVE, DR, SIR-II and OF2. We computed the OF2 algorithm detailed in Section 4.2 and the simulations we made truly argued in favor of its convergence: after 5 iterations the resulting matrix is nearly stable. It was also interesting to compare criterion (12) between the first iteration matrix and the final one. The difference between both was highly significant. Another important point is that OF2 is not as close to DR, SAVE and SIR-II as OF1 is close to SIR. The following simulations highlight this fact and we expect to have a large scope by providing many kinds of models with different parameter settings. We begin this section by providing the results obtained with Gaussian predictors.

Example 3. We consider the three following regression models

$$\text{Model IV: } Y = \tanh\left(\frac{|X^{(1)}|}{2}\right) + 0.1e$$

$$\text{Model V: } Y = 0.4(X^{(1)})^2 + \sqrt{|X^{(2)}|} + 0.2e$$

$$\text{Model VI: } Y = 1.5X^{(1)}X^{(2)}\, e$$

with $e \overset{d}{=} \mathcal{N}(0, 1)$ and $X \overset{d}{=} \mathcal{N}(0, I_p)$. The standardized CS and the CS of these models are equal. For Model IV, the CS is spanned by $(1, 0, \ldots, 0)$, whereas in Model V and VI, it is a two dimensional subspace generated by $(1, 0, \ldots, 0)$ and $(0, 1, 0, \ldots, 0)$. We consider different parameter configurations for which every presented method is in a convenient situation. We compute SAVE, DR, SIR-II and OF2 with (n, p, H) equal to $(100, 6, 5)$, $(500, 10, 5)$ and $(1000, 20, 10)$. For each configuration, 100 simulated random samples have been generated and the resulting boxplots with their averages are presented in Fig. 3.

For all the selected models, OF2 performs better than all other methods. The most significant improvement happens for Model IV in which our method performs better than the others around 99% of the time in the setting $(100, 5, 6)$. When n increased, OF2 was never worse than the others. Note that for $n = 100, 500$, the average error of OF2 is two times smaller than the average error of DR, SAVE or SIR-II. For $n = 1000$ this factor goes to three. The results of the simulations for

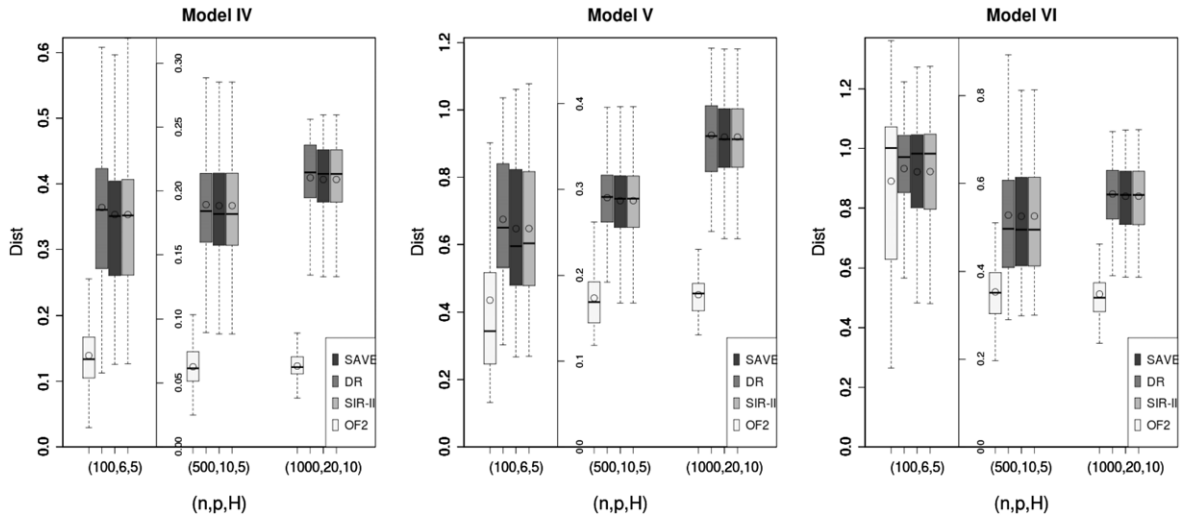

Fig. 3. Plot of the distance error for OF2, DR, SAVE and SIR-II in Example 3.

Model V are really close to Model IV. Model VI is a more complicated one for each method, we have to wait $n = 500$ to notice substantial differences in the distribution of the criterion. In every model, as n increases the improvement of OF2 is substantial. As a consequence and according to the plots in Fig. 1 it seems clear that the asymptotic distribution of the distance error of OF2 has a smaller mean and variance than the other methods. Besides, for the selected models SAVE, DR and SIR-II perform in a similar way and are asymptotically equivalent.

Example 4. To conclude we present the results obtained with non-Gaussian but spherical predictors. Define $X = \rho U$ with U a uniformly distributed vector on the unit sphere of $\mathbb{R}^p$, independent of ρ, a real random variable. Clearly, X has a spherical distribution. Moreover, by taking

$$\rho = \epsilon\,|10 + 0.5N_1| + (1-\epsilon)|30 + 0.5N_2|,$$

with $N_1 \overset{d}{=} \mathcal{N}(0,1)$, $N_2 \overset{d}{=} \mathcal{N}(0,1)$ and $\epsilon \overset{d}{=} \mathcal{B}(\frac{1}{2})$, the distribution of X is far from a normal distribution. We study again Model VI but also the following ones,

$$\text{Model VII: } Y = |X^{(1)}| + \left(\frac{X^{(2)}}{4}\right)^2 + 0.5e$$

$$\text{Model VIb: } Y = X^{(1)}X^{(2)}\,e$$

where $e \overset{d}{=} \mathcal{N}(0,1)$. Model VI has been modified to reduce the signal to noise ratio. The directions to estimate, the parameter configurations and the number of simulated random samples are the same as in the Gaussian case studied previously. Boxplots with their associated averages are presented in Fig. 4.

A general remark regarding Fig. 4 is that the transition from normal to spherical predictors went well for OF2 comparing to other methods. Model IV still reflects the most important improvement of OF2. When n is large, it performs around eight times better than the others. In Model VIb, the accuracy of OF2 deteriorates by changing the distribution of the predictors from Gaussian into spherical. Finally, Model VII provides a standard new situation where the improvement of OF2 is highly significant.

In the development of OF2, Model VI was of particular interest. Whether predictors are normal or spherical, OF2 is highly sensitive to the identification of the CS directions. For $n = 100$ the mean is less than the median, and it is no longer the case for n larger than 100. This marked change in the boxplots is explained by the presence of small outliers in the first situation and large outliers in the second one. Indeed as n is getting larger, OF2 performs better but however the mean is shifted by the presence of outliers that reflects uncommon difficult situations. This results from the eigenvector identification process described in Remark 3. Clearly OF2 relies on the way we identify eigenvectors of M_ψ that belong to E_c. To make that possible, a test of independence between the response and the projected predictors is conducted. Outliers of model VI for n equal to 500 and 1000 are the consequence of a bad eigenvector choice realized by this test. When n is sufficiently large this no longer occurs. When the OF2 algorithm is iterated more than 5 times, it happens only very few times.

7. Concluding remarks

The article introduces the basis of a new methodology for SDR. The introduction of some transformations of the response and the optimization with respect to these transformations were the original ideas of this work and have led us to some

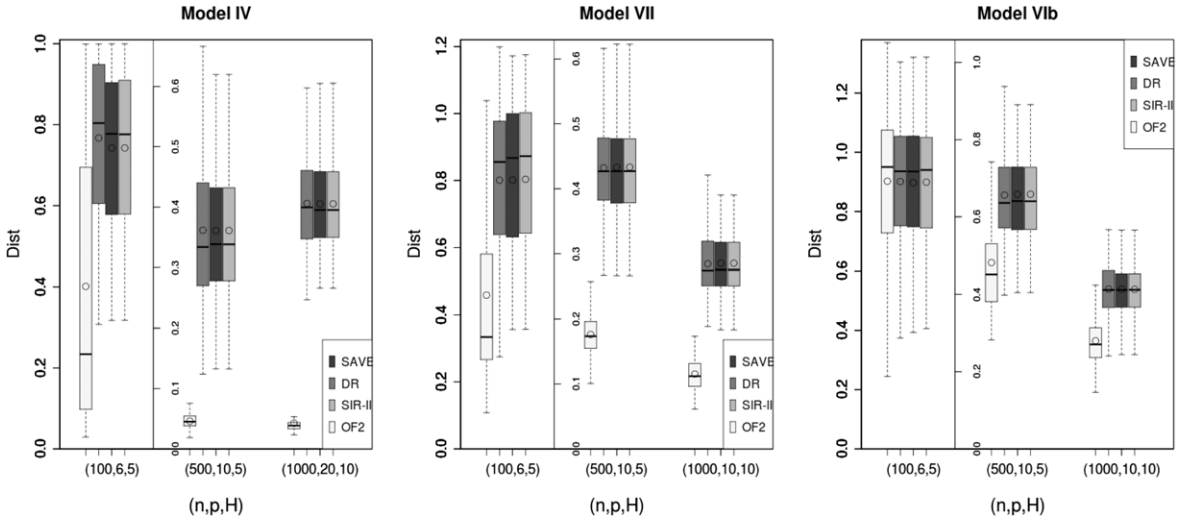

Fig. 4. Plot of the distance error for OF2, SAVE and DR in Example 4.

new methods of investigation in SDR. A surprising point was the high degree of similarity between SIR and OF1. As the simulations pointed out, it could be better to use OF1 when the intra-slice variance is nonconstant. IRE also behaves well in such situations but it has some problems when p is large because of the estimation of a large matrix. Our main contribution relates to order 2 methods, in particular we propose a new class of methods, TF2, that no longer needs the CCV assumption. Moreover, the simulation study sheds light on the high accuracy of OF2 over other order 2 methods. However, one can propose some lines of research that could improve the TF framework.

Regarding the estimation of the dimension, some prospects can be found in the Pearson's chi-squared statistic used in the OF2 algorithm (see Remark 3) to select the eigenvectors that belong to the CS. Clearly, this approach tries to take full advantage of the regression context offered by SDR. Work along this line to estimate the dimension of the CS is in progress and up until now simulations in this sense have provided good results.

Besides, both optimizations OF1 and OF2 do not take into account the estimation error on the variance and the mean of the predictors in the asymptotic decomposition of the criterion (7). This optimization leads to more complicated results that should be validated by simulations.

Finally, in many cases the regression function has different kinds of components, in particular there can be some pathological components for order 1 methods (see Eq. (10)). To handle such cases, one can calculate

$$M = \alpha M_1 + (1-\alpha)M_2,$$

where M_1 and M_2 are matrices of two different SDR methods. A spectral decomposition of M gives a hybrid estimate of the CS. Such ideas were recommended by Gannoun and Saracco [14] and Ye and Weiss [22] proposed a bootstrap method to select the parameter α. This includes the combinations of SIR and SAVE, SIR and pHd, SIR and SIR-II. Besides, it is commonly known that

$$M_{\mathrm{SAVE}} = \mathbb{E}[\mathrm{var}(Z|Y)^2] + M_{\mathrm{SIR}} - I,$$

and that

$$M_{\mathrm{DR}} = \mathbb{E}[\mathbb{E}[(ZZ^T|Y] - I)^2] + M_{\mathrm{SIR}}^2 + \mathrm{tr}(M_{\mathrm{SIR}})M_{\mathrm{SIR}},$$

making SAVE and DR some combinations of SIR and order 2 methods. Therefore SAVE and DR do not only involve order 2 moments of Z, unlike TF2. Moreover TF1 only involves order 1 moments of Z. As a consequence, it seems more realistic to develop hybrid methods based on TF1 and TF2. Especially, the choice of the parameter α could be realized by the optimization of a well chosen criterion as has been done independently to derive OF1 and OF2.

Appendix A. Proofs of the stated results

Proof of Theorem 1. The standardization of the predictors does not change the presentation of this result, hence we present it for X. The proof is divided into three principal parts: we first give a lemma about the intersection of two MDRS, then we apply it to prove the statement of the theorem about the CMS, finally using this last result we conclude the proof for the CS. □

Lemma 11. *If the restriction of X to the ball of $\mathbb{R}^p$ with radius r and center x_0 has a strictly positive density, then the intersection of two MDRS is a MDRS on this ball, i.e.*

$$(\mathbb{E}[Y|X] - \mathbb{E}[Y|RX])\mathbb{1}_{\{X\in B(x_0,r)\}} = 0 \quad a.s.,$$

where R denotes the orthogonal projector onto their intersection.

Proof. We first make the proof for a ball centered at 0, and then we apply it to $X - x_0$. Let E and E' be two MDRS, P and P' their respective orthogonal projectors, and R the orthogonal projector onto $E \cap E'$. Using the definition of a MDRS,

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|PX] = \mathbb{E}[Y|P'X] \quad \text{a.s.}$$

Let $g(PX)$ and $h(P'X)$ denote the last two random variables in the preceding equation. Using that X has a strictly positive density on the unit sphere, we can write

$$g(Px) = h(P'x) \quad \text{a.e. on } B(0, r). \tag{13}$$

Let $\varepsilon > 0$, and φ_k be a unit approximation with compact support $B(0, \varepsilon)$, we define the function $f_k : B(0, r) \to \mathbb{R}$ such that

$$f_k(x) = (g \circ P) * \varphi_k\,(x).$$

Then, we have for all x,

$$f_k(x) = \int g(P(x - y))\varphi_k(y)dy = f_k(Px).$$

Moreover, for all $x \in B(0, r - \varepsilon)$, since in the above integral $x - y \in B(0, r)$, using (13) we derive

$$f_k(x) = (h \circ P') * \varphi_k\,(x),$$

and similarly we obtain $f_k(x) = f_k(P'x)$. Since $f_k(x) = f_k(Px) = f_k(P'x)$, a simple iteration process provides for all $x \in B(0, r - \varepsilon)$,

$$f_k(x) = f_k((PP')^n x).$$

Since f_k is a continuous function and $\lim_{n\to+\infty}(PP')^n = R$, we have

$$f_k(x) = f_k(Rx), \quad x \in B(0, r - \varepsilon).$$

To conclude, the unit approximation theorem gives us the convergence

$$f_k \circ R \xrightarrow{L_1} g \circ P.$$

Thus, from $f_k(RX)$ we can derive a subsequence $f_{n_k}(RX)$ that converges almost surely to $g(PX)$, proving that $\mathbb{E}[Y|X]$ is a function of RX. This completes the first part of the proof.

Now suppose that X has a strictly positive density onto the ball of radius r and center x_0. Define $\widetilde{X} = X - x_0$, it is clear that a MDRS for X is also a MDRS for $\widetilde{X}$ and conversely. Then, since $\widetilde{X}$ is centered in 0, the intersection of two MDRS is still a MDRS for $\widetilde{X}$ and obviously for X. □

Existence of the CMS. Denote by $F \subset \mathbb{R}^p$ the support of the density of X. A first step consists of showing that its interior $\mathring{F}$ can be covered by a countable number of balls included in $\mathring{F}$. Secondly, we apply Lemma 11 to each of this balls to obtain that the intersection of two MDRS on $\mathring{F}$ is a MDRS on $\mathring{F}$. Finally, the uniqueness is shown.

Let $x \in \mathring{F}$, then there exists $r > 0$ such that $B(x, r) \subset \mathring{F}$. It is possible to find a ball, with rational center and radius, included in $B(x, r)$ and containing x. Thus any x of $\mathring{F}$ is contained in a ball with center and radius rational that is included in $\mathring{F}$. In other words, the set A formed by all the balls $B(x_q, r_q) \subset \mathring{F}$, with x_q and r_q rationals, covers $\mathring{F}$. Therefore, by applying Lemma 11, we have for all $B(x_q, r_q) \in A$,

$$|\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]|\mathbb{1}_{\{X\in B(x_q,r_q)\}} = 0 \quad \text{a.s.},$$

since A is a countable set,

$$\sum_{(x_q,r_q)\in A} |\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]|\mathbb{1}_{\{X\in B(x_q,r_q)\}} = 0 \quad \text{a.s.},$$

then,

$$|\mathbb{E}[Y|X] - \mathbb{E}[Y|RX]| \sum_{(x_q,r_q)\in A} \mathbb{1}_{\{X\in B(x_q,r_q)\}} = 0 \quad \text{a.s.}$$

By assumption $\mathbb{P}(X \in \mathring{F}) = 1$, then the right-hand side is almost surely strictly positive, and thus

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|RX] \quad \text{a.s.}$$

Consequently, the intersection of two MDRS is a MDRS. To complete the proof, assume that two MDRS have minimum dimension. Their intersection has at least minimum dimension because it is a MDRS. So they are equal.

Existence of the CS. Using similar arguments about the dimension of vector spaces, we only need to show that the intersection of two DRS is a DRS. Let E and E' be two DRS. By Eqs. (2) and (3), E and E' are also MDRS for the random variables $\mathbb{1}_{Y\in A}$ and X. We have just showed that the intersection of two MDRS is a MDRS. Then for all measurable sets A, $E\cap E'$ is a MDRS for $\mathbb{1}_{Y\in A}$ and X. Equivalently, $E\cap E'$ is a DRS. □

Proof of Theorem 3. Assumption 3 implies that $\{\mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]], k = 1,\dots,p\}$ generates E_c. First, let us show that any vector of this family can be approximated by $\mathbb{E}[Z\phi(Y)]$, where ϕ is a linear combination of functions in Ψ. Let $\varepsilon > 0$ and $k\in\{1,\dots,p\}$, since Ψ is a total family in $L_1(\|Z\|)$, there exists ϕ_k a finite linear combination of functions in Ψ such that

$$\mathbb{E}\left[\|Z\|\,|\phi_k(Y) - \mathbb{E}[Z^{(k)}|Y]|\right] \le \varepsilon,$$

besides, we have

$$\|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]]\| \le \mathbb{E}\left[\|Z\|\ \left|\phi_k(Y) - \mathbb{E}[Z^{(k)}|Y]\right|\right],$$

and therefore,

$$\|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z^{(k)}|Y]]\| \le \varepsilon. \tag{14}$$

Here an important point is that $\mathbb{E}[Z\phi_k(Y)]\in E_c$, it implies that

$$\operatorname{span}\left(\mathbb{E}[Z\phi_k(Y)], k = 1,\dots,p\right) \subset \operatorname{span}(M_{\mathrm{SIR}}). \tag{15}$$

Moreover, (14) and the continuity of the determinant involve that the rank of the set of vectors $\mathbb{E}[Z\phi_k(Y)]$'s is equal to d if ε is small enough. Then, instead of an inclusion (15) becomes an equality and we complete the proof by recalling that each ϕ_k is a linear combination of a finite number of functions in Ψ. □

Proof of Proposition 4. We first calculate the expectation of the limit in law of the sequence $n\operatorname{tr}(Q_c\widehat{P}_c)$ and then we solve the optimization problem. Since

$$n\operatorname{tr}(Q_c\widehat{P}_c) = n\operatorname{tr}(\widehat{\eta}^T Q_c\widehat{\eta}\,(\widehat{\eta}^T\widehat{\eta})^{-1}) = \operatorname{tr}(\sqrt{n}(\widehat{\eta}^T - \eta^T)Q_c\sqrt{n}(\widehat{\eta}-\eta)(\widehat{\eta}^T\widehat{\eta})^{-1}).$$

Slutsky's theorem and the continuity of the operator $\operatorname{tr}(\cdot)$ provide that $n\operatorname{tr}(Q_c\widehat{P}_c)$ converges to $\operatorname{tr}(\delta^T Q_c\delta)$ in distribution, where $\delta\in\mathbb{R}^{p\times d}$ is the limit in law of the sequence $\sqrt{n}(\widehat{\eta}-\eta)$, i.e. a normal vector with mean 0 (we can get ride of the quantity $(\widehat{\eta}^T\widehat{\eta})^{-1}$ because of the constraint and $\widehat{\eta}^T\widehat{\eta}\xrightarrow{\mathbb{P}}\eta^T\eta$). Thus it remains to calculate the expectation of this limit, notice that

$$\mathbb{E}[W_\alpha] = \mathbb{E}\left[\operatorname{tr}(\delta^T Q_c\delta)\right] = \sum_{k=1}^{d}\operatorname{tr}\left(Q_c\mathbb{E}[\delta_k\delta_k^T]\right),$$

where δ_k stands for the limit in law of the sequence $\sqrt{n}(\widehat{\eta}_k-\eta_k)$. Finally, since its variance is equal to $\operatorname{var}(Z\psi_k(Y))$ and using the linearity condition, we find that

$$\mathbb{E}[W_\alpha] = \sum_{k=1}^{d}\mathbb{E}\left[\|Q_cZ\|^2\psi_k(Y)^2\right].$$

Now let us formulate the minimization problem with respect to the matrix α. Using that the $I(h)$ are pairwise disjoint, we have

$$\mathbb{E}[W_\alpha] = \sum_{k=1}^{d}\alpha_k^T\mathbb{E}[\|Q_cZ\|^2\mathbb{1}_Y\mathbb{1}_Y^T]\alpha_k = \operatorname{tr}(\alpha^T D\alpha),$$

and also,

$$\eta^T\eta = \alpha^T C^T C\alpha = (D^{\frac12}\alpha)^T G D^{\frac12}\alpha.$$

From both previous equations, we set out the equivalent minimization problem

$$\min_\alpha \operatorname{tr}\left(\alpha^T D\alpha\right)\quad \text{u.c. } (D^{\frac12}\alpha)^T G D^{\frac12}\alpha = Id,$$

then, from the variable change $U = V^T D^{\frac12}\alpha$ we derive

$$\min_U \operatorname{tr}(U^T U)\quad \text{u.c. } U^T\begin{pmatrix} D_0 & 0\\ 0 & 0\end{pmatrix}U = Id.$$

By writing $U^T = (U_1^T, U_2^T)$ we notice that there is no constraint on U_2, which implies that $U_2 = 0$. Consequently, it remains to solve

$$\min_{U_1} \operatorname{tr}(U_1 U_1^T) \quad \text{u.c. } U_1 U_1^T = D_0^{-1},$$

where $U_1 \in \mathbb{R}^{d\times d}$, and where the quantity to minimize is fixed by the constraint. Then, a solution is $U_1 = D_0^{-\frac{1}{2}} H$ where H is any orthogonal matrix. Hence, the solution of the minimization problem is

$$\alpha = D^{-\frac{1}{2}} V U = D^{-\frac{1}{2}} V_1 D_0^{-\frac{1}{2}} H. \quad \square$$

Proof of Lemma 5. Let us begin in the easiest way : (2) $\Rightarrow$ (1). Let H be any orthonormal matrix as described in (1). Because $HQ_cH^T = I - HP_cH^T = Q_c$, by multiplying (2) on the left side by H and on the right side by H^T, we find that

$$\operatorname{var}(HZ|P_cZ) = \lambda_\omega^* Q_c = \operatorname{var}(Z|P_cZ).$$

The other way is based on a good choice of the matrix H. Let γ be a unit vector of $E_c^\perp$, and define $H = I - 2\gamma\gamma^T$. Clearly, H is symmetric and satisfies to the requirement of (1). So that, we have

$$\operatorname{var}(Z|P_cZ) = (I - 2\gamma\gamma^T)\operatorname{var}(Z|P_cZ)(I - 2\gamma\gamma^T),$$

developing the right hand side, it follows that

$$\operatorname{var}(Z|P_cZ)\gamma\gamma^T = 2\operatorname{var}(\gamma^T Z|P_cZ)\gamma\gamma^T - \gamma\gamma^T\operatorname{var}(Z|P_cZ),$$

and finally, multiplying by γ on the right, we find

$$\operatorname{var}(Z|P_cZ)\gamma = \operatorname{var}(\gamma^T Z|P_cZ)\gamma. \tag{16}$$

Therefore, any $\gamma \in E_c^\perp$ is an eigenvector of the matrix $\operatorname{var}(Z|P_cZ)$ and thus, $E_c^\perp$ is included in an eigenspace of this matrix. Denote by λ_ω^* the eigenvalue associated with $E_c^\perp$. Since the columns of Q_c are vectors of $E_c^\perp$, we have

$$\operatorname{var}(Z|P_cZ)Q_c = \lambda_\omega^* Q_c,$$

which implies that

$$\operatorname{var}(Z|P_cZ) = \operatorname{var}(Q_cZ|P_cZ) = \lambda_\omega^* Q_c,$$

and (1) $\Rightarrow$ (2) is completed.

The value of λ_ω^* can be given by Eq. (16). Under the linearity condition we have for every unit vector $\gamma \in E_c^\perp$,

$$\lambda_\omega^* = \operatorname{var}(\gamma^T Z|P_cZ) = \mathbb{E}[(\gamma^T Z)^2|P_cZ],$$

and hence it suffices to take $\gamma = \frac{1}{\sqrt{p-d}}\sum_{k=1}^{p-d}\gamma_k$ where $(\gamma_1, \ldots, \gamma_{p-d})$ is an orthonormal basis of $E_c^\perp$, to obtain

$$\lambda_\omega^* = \frac{1}{p-d}\mathbb{E}\left[\|Q_cZ\|^2|P_cZ\right]. \quad \square$$

Proof of Theorem 6. To make a complete proof, we need to show that all the vectors in $E_c^\perp$ are eigenvectors of the symmetric matrix $M_\psi - \lambda_\psi^* I$ associated with the eigenvalue 0. The existence of the CS ensures that

$$M_\psi - \lambda_\psi^* I = \mathbb{E}[(\mathbb{E}[ZZ^T|P_cZ] - \lambda_\omega^* I)\psi(Y)],$$

besides, thanks to the linearity condition and DCV, we have

$$\mathbb{E}[ZZ^T|P_cZ] = \lambda_\omega^* Q_c + P_cZZ^TP_c.$$

Thus, for any $\gamma \in E_c^\perp$ we have $(M_\psi - \lambda_\psi^* I)\gamma = 0$ and the proof is completed. $\square$

Proof of Theorem 7. The proof relies on Lemmas B.4 and B.5. Both are results about vector spaces of non-invertible matrices. For clarity and since it does not deal directly with the subject of the paper, we state and prove these lemmas in Appendix B.

Let Ψ be a total countable family in $L_1(\|Z\|^2)$, Theorem 6 indicates that $E_c^\perp \subset E_\psi^\perp$ for any $\psi \in \Psi$. Then it suffices to show that there exists ψ a finite linear combination of functions in Ψ such that $\dim(E_\psi) = \operatorname{rank}(M_\psi - \lambda_\psi^* I) = d$. In the basis (P_1, P_2), where P_1 and P_2 are respectively bases of E_c and $E_c^\perp$, the matrix $M_\psi - \lambda_\psi^* I$ can be written as

$$\begin{pmatrix} N_\psi & 0 \\ 0 & 0 \end{pmatrix},$$

with $N_\psi = P_1^T(M_\psi - \lambda_\psi^*)P_1$. Notice that the space

$$\mathcal{M} = \left\{ N_\psi,\ \psi = \sum_h \alpha_h \psi_h \right\},$$

is a vector space of symmetric matrices with dimension $d \times d$. In the basis (P_1, P_2), Assumption 5 becomes

$$\forall \eta \in \mathbb{R}^d, \quad \mathbb{P}(\eta^T N_Y \eta = 0) < 1,$$

with $N_Y = P_1^T(M_Y - \lambda_Y^*)P_1$. Clearly, this implies that

$$\forall \eta \in \mathbb{R}^d, \exists \psi, \quad \eta^T N_\psi \eta \neq 0, \tag{17}$$

and because Ψ is a total family in $L_1(\|Z\|^2)$, the function ψ in the previous equation could be a finite linear combination of functions in Ψ and then $N_\psi \in \mathcal{M}$. Thus to conclude the proof, one can notice that given a vector subspace $\mathcal{M} \subset \mathbb{R}^{d\times d}$ of symmetric matrices, if (17) holds, then there exists an invertible matrix in $\mathcal{M}$. This assertion is true because it is the contrapositive of the statement of Lemma B.5. □

Proof of Corollary 8. From Theorem 7 we have $E_\psi = E_c$ where $\psi = \sum_{h=1}^H \alpha_h \psi_h$. Hence, we need to show that $E_\psi \subset \oplus E_{\psi_h}$ since the other inclusion is trivial. Suppose that there exists $\eta \in E_\psi$ with norm 1 such that $\eta \perp \oplus E_{\psi_h}$. Then by definition, for every $h = 1, \ldots, H$, we have

$$M_{\psi_h}\eta = \lambda_{\psi_h}^* \eta,$$

and we can obtain

$$M_\psi \eta = \sum_{h=1}^H \alpha_h \lambda_{\psi_h}^* \eta = \lambda_\psi^* \eta,$$

which is impossible because $\eta \in E_\psi$. □

Proof of Proposition 9. We have

$$\begin{aligned} Q_c\widehat{P}_c &= Q_c(\widehat{P}_c - P_c) \\ &= Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1} - (Iz - M_\psi)^{-1} dz \\ &= Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1} dz, \end{aligned}$$

and then, we can obtain

$$\begin{aligned} Q_c\widehat{P}_c = {} & Q_c \oint_{\mathcal{C}} (Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1} dz \\ & + Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1} dz. \end{aligned}$$

Consider the trace of the first term in the above equation, since Q_c and $(Iz - M_\psi)^{-1}$ commute we have

$$\operatorname{tr}\left(Q_c \oint_{\mathcal{C}} (Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1} dz \right) = \operatorname{tr}\left((M_\psi - \widehat{M}_\psi) \oint_{\mathcal{C}} Q_c (Iz - M_\psi)^{-2} dz \right).$$

Besides, it is clear that

$$Q_c(Iz - M_\psi)^{-1} = \frac{Q_c}{(z - \lambda_\psi^*)}, \tag{18}$$

and recalling that λ_ψ^* is outside $\mathcal{C}$, we have $\oint_{\mathcal{C}} \frac{1}{(z-\lambda_\psi^*)^{-2}} dz = 0$ and we get

$$\operatorname{tr}\left(Q_c\widehat{P}_c\right) = \operatorname{tr}\left(Q_c \oint_{\mathcal{C}} (Iz - \widehat{M}_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1}(M_\psi - \widehat{M}_\psi)(Iz - M_\psi)^{-1} dz \right).$$

Denote by Δ the limit in law of $\sqrt{n}(\widehat{M}_\psi - M_\psi)$, since $\widehat{M}$ goes to M in probability, Slutsky's Theorem implies the convergence $n \operatorname{tr}(Q_c\widehat{P}_c) \xrightarrow{d} W_\psi$ with

$$W_\psi = \operatorname{tr}\left(Q_c \oint_{\mathcal{C}} (Iz - M_\psi)^{-1}\Delta(Iz - M_\psi)^{-1}\Delta(Iz - M_\psi)^{-1} dz \right).$$

Now we use Eq. (18) to obtain

$$W_\psi = \operatorname{tr}\left(Q_c \Delta \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz \Delta Q_c \right), \tag{19}$$

where the above integral can be calculated in the following way. Splitting it into two terms and using (18), we have

$$\begin{aligned}\oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz &= \oint_{\mathcal{C}} \frac{P_c(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz + \oint_{\mathcal{C}} \frac{Q_c(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz \\ &= \oint_{\mathcal{C}} \frac{P_c(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz + Q_c \oint_{\mathcal{C}} \frac{1}{(z-\lambda_\psi^*)^3} dz.\end{aligned}$$

It is not difficult to show that the last term in the previous equation equals 0. Regarding the first term, since for every $k \in \{1, \dots, d\}$ we have

$$P_c \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz \eta_k = \eta_k \oint_{\mathcal{C}} \frac{(z - \lambda_\psi(\eta_k))^{-1}}{(z-\lambda_\psi^*)^2} dz = \frac{\eta_k}{(\lambda_\psi(\eta_k) - \lambda_\psi^*)^2} = P_c(P_c M_\psi - I\lambda_\psi^*)^{-2}\eta_k,$$

and since all the vectors in $E_c^\perp$ belong to the kernel of this matrix, we get

$$P_c \oint_{\mathcal{C}} \frac{(Iz - M_\psi)^{-1}}{(z-\lambda_\psi^*)^2} dz = P_c(P_c M_\psi - I\lambda_\psi^*)^{-2}.$$

Injecting it in (19), we obtain

$$W_\psi = \operatorname{tr}\left(\Delta Q_c \Delta P_c (P_c M_\psi - I\lambda_\psi^*)^{-2} \right),$$

and it remains to calculate its expectation. The linearity condition implies that $Q_c M_\psi P_c = 0$, and we have

$$\mathbb{E}[\Delta Q_c \Delta P_c] = \lim_{n \to +\infty} n\mathbb{E}\left[\widehat{M}_\psi Q_c \widehat{M}_\psi P_c \right] = \mathbb{E}[ZZ^T P_c \|Q_c Z\|^2 \psi(Y)^2],$$

which completes the proof. □

Proof of Theorem 10. The proof involves a result in [4], stated in Appendix B as Theorem B.3.

By applying Theorem B.3 to the matrix $\widehat{C}\widehat{\alpha}$, one can notice that the asymptotic distribution of $\widehat{\Lambda}_{\mathrm{TF1}}$ depends only on the variance of the asymptotic law of

$$\sqrt{n}\operatorname{vec}(U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0).$$

Let W be a random vector following this distribution. By the linearity condition, we have

$$U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0 = U_0^T\widehat{C}\widehat{\alpha}V_0 = U_0^T\widehat{\Sigma}^{-1/2}(\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C} - \widehat{\Sigma}^{-\frac{1}{2}}C)\widehat{\alpha}V_0.$$

Since $C\alpha V_0 = 0$, $\widehat{\alpha} \xrightarrow{\mathbb{P}} \alpha$ and $\widehat{\Sigma} \xrightarrow{\mathbb{P}} \Sigma$, by Slutsky's theorem W has the same law as the asymptotic distribution of

$$\sqrt{n}\operatorname{vec}(U_0^T\Sigma^{-1/2}\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C}\alpha V_0).$$

By the linearity condition $U_0\Sigma^{-1/2}X_i = U_0\Sigma^{-1/2}(X_i - \mathbb{E}[X]) = U_0 Z_i$, and one can obtain

$$U_0^T\Sigma^{-1/2}\widehat{\Sigma}^{-\frac{1}{2}}\widehat{C}\alpha V_0 = U_0^T(\overline{Z\Psi_H^T(Y)} - \overline{Z}\,\overline{\Psi_H^T(Y)})\alpha V_0.$$

We notice that

$$\sqrt{n}(\overline{Z\Psi_H^T(Y)} - \overline{Z}\,\overline{\Psi_H^T(Y)}) = \sqrt{n}\left(\overline{Z(\Psi_H^T(Y) - \mathbb{E}[\Psi_H^T(Y)])}\right) + o_{\mathbb{P}}(1),$$

and we provide the decomposition

$$\sqrt{n}\operatorname{vec}(U_0^T(\widehat{C}\widehat{\alpha} - C\alpha)V_0) = (V_0^T\alpha^T \otimes U_0^T)\sqrt{n}\operatorname{vec}\left(\overline{Z\Phi(Y)}\right) + o_{\mathbb{P}}(1),$$

with the notation $\Phi(Y) = \Psi_H(Y) - \mathbb{E}[\Psi_H(Y)]$. By the central limit theorem, we get

$$\frac{1}{\sqrt{n}}\sum_{i=1}^n \operatorname{vec}(Z_i\Phi(Y_i)^T) \xrightarrow{d} \mathcal{N}\left(0, \operatorname{var}(\Phi(Y)\otimes Z)\right).$$

Clearly, using the linearity condition we have

$$\operatorname{var}(W) = (V_0^T\alpha^T \otimes I)\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^T ZZ^T U_0)](\alpha V_0 \otimes I).$$

Under DCV one can get

$$\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^T ZZ^T U_0)] = \mathbb{E}[(p-d)^{-1}\|Q_c Z\|^2 \Phi(Y)\Phi(Y)^T \otimes I_{p-d}],$$

under CCV one can obtain

$$\mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes (U_0^T ZZ^T U_0)] = \mathbb{E}[\Phi(Y)\Phi(Y)^T \otimes I_{p-d}],$$

and the conclusion follows. □

Appendix B. Few results

Proposition B.1 (*[1, Theorem 4.1.4, p. 48]*). *Let Z be a random vector of $\mathbb{R}^p$ ($p \geq 2$) with a finite second order moment. If Z is spherical and if $\mathrm{var}(Z|PZ) = const$ for some orthogonal projector P, then Z is normal and conversely.*

Theorem B.2 (*[5]*). *Let $p \in [0, +\infty[$, μ a Borel probability measure on $[0, 1]$, and $f_n : [0, 1] \to \mathbb{R}$ a family of bounded measurable functions that separates the points:*

$$\forall x, y \in [0, 1], x \neq y, \exists n \in \mathbb{N} \quad \text{such that } f_n(x) \neq f_n(y).$$

Then the algebra spanned by the functions f_n's and the constants is dense in $L_p([0, 1], \mu)$.

Theorem B.3 (*[4]*). *Assume $\mathrm{rank}(M) = d$ and that $\sqrt{n}\,\mathrm{vec}(\widehat{M} - M) \xrightarrow{d} \mathcal{N}(0, \Gamma)$. Then*

$$\widehat{\Lambda} \xrightarrow{d} \sum_{k=1}^{s} \omega_k X_k^2,$$

where the X_k's are independent standard normal random variables and the ω_k's are the ordered eigenvalues of $(V^T \otimes U^T)\Gamma(V \otimes U)$, with $s = \min(\mathrm{rank}(\Gamma), (p-d)(H-d))$ and U and V are respectively basis of the left and right singular spaces of M associated with the singular value 0.

The following lemma deals with vector space structure and rank-deficient matrices. We refer to [11, Proposition 3], for a more general approach. In particular, this lemma implies Lemma B.5 which has a central place in the proof of Theorem 7.

Lemma B.4. *Let $M, N \in \mathbb{R}^{d\times d}$ and $\alpha_0 > 0$. If $\mathrm{rank}(N + \alpha M) \leq \mathrm{rank}(N)$ for all $\alpha \leq \alpha_0$, then we have*

$$M \ker(N) \subset \mathrm{Im}(N).$$

Proof. Denote by P_α the characteristic polynomial of $N+\alpha M$ and define $r_\alpha = \mathrm{rank}(N+\alpha M)$ and $k_\alpha = \dim(\ker(N+\alpha M)) = d - r_\alpha$. Because of the continuity of the determinant, the coefficients of P_α converge to the coefficients of P_0, then P_α converges uniformly to P_0 on every compact. By the definition of k_0, P_0 is such that

$$P_0(x) = x^{k_0} Q_0(x) \quad \text{with } Q_0(0) \neq 0.$$

Now we use the uniform convergence. For α small enough we have $P_\alpha^{(k_0)}(0) \neq 0$, and this gives the upper bound $k_\alpha \leq k_0$. Using the assumption we obtain $k_0 = k_\alpha$. Therefore, for some α_0, we have

$$Q_\alpha(0) \neq 0, \quad \alpha \leq \alpha_0.$$

Clearly, there exists a contour $\mathcal{C}$ such that none of the nonzero eigenvalues of $N+\alpha M$ belong to $\mathcal{C}$, $\alpha \leq \alpha_0$. Using the residue theorem, we can express the orthogonal projectors Π_0 and Π_α on the kernel of the matrices N and $N + \alpha M$ as follows,

$$\Pi_0 = \oint_{\mathcal{C}} (N - zI)^{-1} dz, \quad \text{and} \quad \Pi_\alpha = \oint_{\mathcal{C}} (N + \alpha M - zI)^{-1} dz,$$

and one can get

$$\Pi_0 - \Pi_\alpha = \alpha \oint_{\mathcal{C}} (N - zI)^{-1} M (N + \alpha M - zI)^{-1} dz.$$

Because as α goes to 0, none of the eigenvalues of N and $N + \alpha M$ crosses $\mathcal{C}$, the integral converges and then we derive that $\lim_{\alpha\to 0} \Pi_\alpha = \Pi_0$. Besides, we have

$$(N + \alpha M)\Pi_\alpha = 0, \quad \text{and} \quad N\Pi_0 = 0,$$

then we get $N(\Pi_0 - \Pi_\alpha) = \alpha M \Pi_\alpha$, and we obtain

$$\mathrm{Im}(M\Pi_\alpha) \subset \mathrm{Im}(N).$$

We conclude the proof using the continuity of Π_α. □

Lemma B.5. *Let $\mathcal{M} \subset \mathbb{R}^{d\times d}$ be a vector space of non-invertible symmetric matrices. We have*

$$\exists u \in \mathbb{R}^d, \forall M \in \mathcal{M}, \quad u^T M u = 0.$$

Proof. Since $\mathcal{M}$ is a vector space, we can apply Lemma B.4 with N a matrix of maximal rank in $\mathcal{M}$ and any $M \in \mathcal{M}$. Then, for every $u \in \ker(N)$, there exists $y \in \mathbb{R}^d$ such that

$$Mu = Ny.$$

Because N is symmetric, by multiplying the left-hand side by u^T, we obtain $u^T M u = 0$. □

References

[1] Włodzimierz Bryc, The Normal Distribution, in: Lecture Notes in Statistics, vol. 100, Springer-Verlag, New York, 1995, Characterizations with applications.
[2] Efstathia Bura, Dimension reduction via parametric inverse regression, in: L_1-Statistical Procedures and Related Topics (Neuchatel, 1997), in: IMS Lecture Notes Monogr. Ser., vol. 31, Inst. Math. Statist., Hayward, CA, 1997, pp. 215–228.
[3] Efstathia Bura, R. Dennis Cook, Extending sliced inverse regression: the weighted chi-squared test, J. Amer. Statist. Assoc. 96 (455) (2001) 996–1003.
[4] E. Bura, J. Yang, Dimension estimation in sufficient dimension reduction: a unifying approach, J. Multivariate Anal. 102 (1) (2011) 130–142.
[5] Y. Coudène, Une version mesurable du théorème de Stone–Weierstrass, Gaz. Math. (91) (2002) 10–17.
[6] Arnak S. Dalalyan, Anatoly Juditsky, Vladimir Spokoiny, A new algorithm for estimating the effective dimension-reduction subspace, J. Mach. Learn. Res. 9 (2008) 1648–1678.
[7] R. Dennis Cook, Regression Graphics, in: Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1998.
[8] R. Dennis Cook, Bing Li, Dimension reduction for conditional mean in regression, Ann. Statist. 30 (2) (2002) 455–474.
[9] R. Dennis Cook, Liqiang Ni, Sufficient dimension reduction via inverse regression: a minimum discrepancy approach, J. Amer. Statist. Assoc. 100 (470) (2005) 410–428.
[10] R. Dennis Cook, Sanford Weisberg, Discussion of "sliced inverse regression for dimension reduction", J. Amer. Statist. Assoc. (1991) 28–33.
[11] Jan Draisma, Small maximal spaces of non-invertible matrices, Bull. London Math. Soc. 38 (5) (2006) 764–776.
[12] Morris L. Eaton, A characterization of spherical distributions, J. Multivariate Anal. 20 (2) (1986) 272–276.
[13] Morris L. Eaton, David E. Tyler, On Wielandt's inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix, Ann. Statist. 19 (1) (1991) 260–271.
[14] Ali Gannoun, Jérôme Saracco, An asymptotic theory for SIR_α method, Statist. Sinica 13 (2) (2003) 297–310.
[15] Marian Hristache, Anatoli Juditsky, Jörg Polzehl, Vladimir Spokoiny, Structure adaptive approach for dimension reduction, Ann. Statist. 29 (6) (2001) 1537–1566.
[16] Ker-Chau Li, Sliced inverse regression for dimension reduction, J. Amer. Statist. Assoc. 86 (414) (1991) 316–342.
[17] Ker-Chau Li, On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma, J. Amer. Statist. Assoc. 87 (420) (1992) 1025–1039.
[18] Bing Li, Yuexiao Dong, Dimension reduction for nonelliptically distributed predictors, Ann. Statist. 37 (3) (2009) 1272–1298.
[19] Bing Li, Shaoli Wang, On directional regression for dimension reduction, J. Amer. Statist. Assoc. 102 (479) (2007) 997–1008.
[20] Bing Li, Hongyuan Zha, Francesca Chiaromonte, Contour regression: a general approach to dimension reduction, Ann. Statist. 33 (4) (2005) 1580–1616.
[21] Yingcun Xia, Howell Tong, W.K. Li, Li-Xing Zhu, An adaptive estimation of dimension reduction space, J. R. Stat. Soc. Ser. B Stat. Methodol. 64 (3) (2002) 363–410.
[22] Zhishen Ye, Robert E. Weiss, Using the bootstrap to select one of a new class of dimension reduction methods, J. Amer. Statist. Assoc. 98 (464) (2003) 968–979.
[23] Xiangrong Yin, R. Dennis Cook, Dimension reduction for the conditional kth moment in regression, J. R. Statist. Soc. Ser. B Statist. Methodol. 64 (2) (2002) 159–175.
[24] Li-Xing Zhu, Kai-Tai Fang, Asymptotics for kernel estimate of sliced inverse regression, Ann. Statist. 24 (3) (1996) 1053–1068.