Asymptotic optimality for sliced inverse regression

Portier, François
IRMAR, University of Rennes1
Campus de Beaulieu
Rennes 35042, France
E-mail: francois.portier@univ-rennes1.fr

Delyon, Bernard
IRMAR, University of Rennes1
Campus de Beaulieu
Rennes 35042, France
E-mail: bernard.delyon@univ-rennes1.fr

1 Introduction

Dimension reduction in regression aims at improving poor convergence rates derived from the non-parametric 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* (SDR) 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 $X = (X_1, ..., X_p)^T$ by a number smaller than p of linear combination of the predictors. The new covariate vector has the form PX, where P can be chosen as an orthogonal projection 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 preservation of the information carried by X about Y through the projection on E. In [5] and [2] a *dimension reduction subspace* (DRS) is defined by the conditional independence property

$$(1) \quad Y \perp\!\!\!\perp X \mid P_c X,$$

where P_c is the orthogonal projection on a DRS. With 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

$$(2) \quad \mathbb{P}(Y \in A|X) = \mathbb{P}(Y \in A|P_c X),$$

for any measurable set A. Moreover under some additional condition (see [2]), the intersection of all the DRS is itself a DRS. Consequently, there exists a unique DRS with minimal dimension and we call it the *central subspace* (CS) [2]. In this article the CS is noted E_c and we assume its existence.

The present work deals with a part of the literature based on the principle of inverse regression; i.e. instead of studdying the regression curve which implies high dimensional estimation problems, the study is focused on the inverse regression curve $\mathbb{E}[X|Y = y]$ (order 1) or the inverse variance curve $\mathrm{var}(X|Y = y)$ (order 2). Here we are only concerned with the order 1 moment based methods. These methods include *sliced inverse regression* (SIR) [5], *kernel inverse regression* (KIR) [6], *parametric inverse regression* (PIR) [1], and *inverse regression estimator* (IRE) [3]. For many different aspects not detailed here, SIR is the indisputable leader of the previous enumerate and we refer to [3] for a full background about inverse regression.

As quoted before, there exists a large range of methods aiming at the estimation of the CS. By introducing the order 1 test function methodology (TF1), we try to propose a general point of view about SDR. The original basic idea of TF1 is to investigate the dependence between X and Y

by introducing nonlinear transformations of Y, and inferring about the CS through their covariances with X. Hence, the CS is obtained by inspection of the range of $\mathbb{E}[X\psi(Y)]$, where ψ varies in a well chosen family of functions. Methods deriving from TF1 provide an exhaustive estimation of the CS under the same conditions than SIR. Moreover, an asymptotic variance analysis leads us to the optimal transformation of Y for the estimation of the CS.

This article is organized as follows. In section 2, we introduce TF1 and the conditions for its exhaustiveness. The choice of the optimal transformation of the response is detailed in section 3 in which a plug-in method is proposed.

2 The test function methodology

To explain our next results in a simple way, we introduce the standardized covariate $Z = \Sigma^{-\frac{1}{2}}(X - \mathbb{E}[X])$ with $\Sigma = \text{var}(X)$. Hence we define the standardized central subspace equal to $\Sigma^{\frac{1}{2}}E_c$. Since there is no ambiguity we always note it E_c. We denote by P_c the orthogonal projection on E_c and we note $\dim(E_c) = d$ and $Q_c = I - P_c$. For any matrix M, we note $\text{span}(M)$ the space generated by the columns of M.

2.1 Basic ideas

Model (1) implies that all the information detained by Z about Y is carried by P_cZ. To find E_c, as pointed out by [5] and explained in many articles on the subject, a natural idea is to focus on the inverse regression curve $\mathbb{E}[Z|Y]$. Actually, if (1) holds, we can write the inverse regression curve as $\mathbb{E}[\mathbb{E}[Z|P_cZ]|Y]$. If in addition, $\mathbb{E}[Z|P_cZ] \in E_c$, then $\mathbb{E}[Z|Y]$ is with probability 1 a vector of E_c. The previous idea is the cornerstone of many dimension reduction methods and TF1 is also based on it. SIR consists in estimating the matrix $M_{SIR} = \mathbb{E}[\mathbb{E}[Z|Y]\mathbb{E}[Z|Y]]$ which column space is included in the CS, whereas TF1 general approach is interested in vector families of the kind $\mathbb{E}[Z\psi_1(Y)], ..., \mathbb{E}[Z\psi_q(Y)]$ for some set of measurable functions $\psi_k : \mathbb{R} \to \mathbb{R}$. We need an assumption frequently used in dimension reduction, called the linearity condition.

Assumption 1. *(linearity condition)*

$$\mathbb{E}[Q_cZ|P_cZ] = 0 \quad a.s.$$

Theorem 1. *Assume that Z satisfies Assumption 1 and has a finite first moment. Then, for every measurable function $\psi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[Z\psi(Y)] < \infty$, we have*

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

Proof. Thanks to the existence of the central subspace, $\mathbb{E}[Z\psi(Y)] = \mathbb{E}\left[\mathbb{E}[Z|P_cZ]\psi(Y)\right]$, and thanks to the linearity condition, $Q_c\mathbb{E}[Z\psi(Y)] = 0$. □

The previous theorem is not really new. Yet, it makes a simple link between TF1 and the CS by providing a vector in E_c for every measurable function. Without additional assumption none of the spaces generated by SIR and TF1 cover the entire CS.

2.2 Covering the central subspace

In this article, two kinds of conditions can be distinguished. Those that allow such a characterization of the CS, and those that guarantee to cover the entire CS. As a consequence of Theorem 1, spaces

generated by $(\mathbb{E}[Z\psi_1], ..., \mathbb{E}[Z\psi_q])$ are included in E_c. Our goal is to obtain the converse inclusion. Because TF1 is an extending 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 2. *For every nonzero vectors $\eta \in E_c$, $\mathbb{E}[\eta^T Z|Y]$ has a nonzero variance.*

Lemma 1. *If Z satisfies Assumption 1 and has a finite second moment, then Assumption 2 implies that* $\text{span}(M_{SIR}) = E_c$ *and conversely.*

Proof. Under the linearity condition, $\text{span}(M_{SIR}) = E_c$ is equivalent to $\eta^T M_{SIR}\ \eta > 0$ for every $\eta \in E_c$. □

We now extend Lemma 1 to TF1. To state the following theorem, we introduce the function space $L_1(\theta(y)\mu(dy))$ defined as

$$L_1(\theta(y)\mu(dy)) = \{u : \mathbb{R} \to \mathbb{R}; \int_{\mathbb{R}} |u(y)|\theta(y)\mu(dy) < +\infty\},$$

where $\theta : \mathbb{R} \to \mathbb{R}_+$ and μ a real measure.

Theorem 2. *Assume that Z and Y satisfy Assumptions 1 and 2. Assume also that Z has a finite second moment. If Ψ is a total countable family in the space $L_1(\mathbb{E}[\|Z\||Y = y]P_Y(dy))$, then we can extract a finite subset Ψ_H of Ψ such that* $\text{span}\left(\mathbb{E}[Z\psi(Y)],\ \psi \in \Psi_H\right) = E_c$.

Proof. Lemma 1 provides that $\{\mathbb{E}[Z\mathbb{E}[Z_k|Y]],\ k = 1, ..., p\}$ is a generator of 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 $k \in \{1, ..., p\}$. Since Ψ is a total family in $L_1(\mathbb{E}[\|Z\||Y = y]P_Y(dy))$, there exists ϕ_k a finite linear combination of functions in Ψ such that,

$$\mathbb{E}\left[\mathbb{E}[\|Z\||Y]\ |\phi_k(Y) - \mathbb{E}[Z_k|Y]|\right] \leq \varepsilon,$$

besides, we have

$$\begin{aligned} \|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z_k|Y]]\| &= \|\ \mathbb{E}\left[\mathbb{E}[Z|Y]\ (\phi_k(Y) - \mathbb{E}[Z_k|Y])\right]\| \\ &\leq \mathbb{E}\left[\mathbb{E}[\|Z\|\ |Y]\ |\phi_k(Y) - \mathbb{E}[Z_k|Y]|\ \right], \end{aligned}$$

and therefore,

$$\|\mathbb{E}[Z\phi_k(Y)] - \mathbb{E}[Z\mathbb{E}[Z_k|Y]]\| \leq \varepsilon. \quad (3)$$

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

$$\text{Span}\left(\mathbb{E}[Z\phi_k(Y)],\ k = 1, ..., p\right) \subset \text{span}(M_{SIR}), \quad (4)$$

Moreover, (3) and the continuity of the determinant involve that the rank of the set of vectors $\mathbb{E}[Z\phi_k(Y)]$ is equal to d if ε is small enough. Then, instead of an inclusion (4) become 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. □

Theorem 2 assumes that the family is total. Some mild conditions can be found in [4]. Let us recall their main result.

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

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

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

Accordingly, we can apply Theorem 2 with any family of functions that separates the points, for example polynomials, complex exponentials or indicator functions. To make possible a simple use of this theorem we need to recall this result. If $u = (u_1, ..., u_H)$ is a $\mathbb{R}^p$ vector family, then $\text{span}(uu^T) = \text{span}(u)$. Thus, if we denote by $\psi_1, ..., \psi_H$ 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 to the functions $\psi_1, ..., \psi_H$ defined as

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

Especially, the eigenvectors associated to a nonzero eigenvalue of any order 1 test function matrix span the central subspace.

3 Choice of the test function for asymptotic optimality

Theorem 2 implies that the subspace E_c can be covered in totality by the family of vectors $\{\mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}], h = 1, ..., H\}$. Actually, it is possible to extract d orthogonal vectors living in the space spanned by this family, and then it provides us a basis of the central subspace. This procedure is realized by SIR. Nevertheless, the issue here is somewhat more complicated, we want to find d orthogonal vectors that have the minimal asymptotic mean squared error for the estimation of the projection P_c. We define

$$\text{MSE} = \mathbb{E}\left[\|P_c - \widehat{P}_n\|^2\right], \tag{5}$$

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

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

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

$$\begin{aligned} \text{MSE} &= \mathbb{E}[\text{tr}(P - \widehat{P}_n)^2] \\ &= d + \mathbb{E}[\widehat{d} - 2\text{tr}((I - Q_c)\widehat{P}_n)] \\ &= \mathbb{E}[d - \widehat{d}] + 2\mathbb{E}[\text{tr}(Q_c\widehat{P}_n)], \end{aligned} \tag{6}$$

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

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

under the constraint of orthogonality of the family $(\eta_k)_{1\leq k\leq d}$. For a more comprehensive approach, we choose to minimize the expectation of the limit in distribution, instead of the limit of the expectation when n goes to infinity, of the sequence $n\text{tr}(Q_c\widehat{P}_n)$. To set out clearly the next proposition, let us introduce some notations. Define the matrices

$$\begin{aligned} C = (C_1, ..., C_H) \qquad &\text{with} \quad C_h = \mathbb{E}[Z\mathbb{1}_{\{Y\in I(h)\}}], \\ D = \text{diag}d_h \qquad &\text{with} \quad d_h = \left(\mathbb{E}\left[\|Q_cZ\|^2\mathbb{1}_{\{Y\in I(h)\}}\right]\right), \end{aligned}$$

and

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

The matrix G is the Gram matrix of the vector family $(C_h/\sqrt{d_h})_{1\leq h\leq H}$, Theorem 2 ensure that its rank is equal to d. Besides, G is diagonalisable and so we define $P = (P_1 P_2) \in \mathbb{R}^{p\times(d+(p-d))}$ such that

$$P^T G P = \begin{pmatrix} D_0 & 0 \\ 0 & 0 \end{pmatrix},$$

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

Proposition 1. *The random variable $n tr(QP_n)$ has a limit in law W_α as $n \to \infty$. The minimization problem*

$$\min_{\alpha} \mathbb{E}\left[W_\alpha\right] \quad u.c. \quad \eta^T \eta = Id, \tag{7}$$

has a unique solution, up to orthogonal transformations, given by $\alpha = D^{-\frac{1}{2}} P_1 D_0^{-\frac{1}{2}}$.

Proof. We first calculate the expectation of the limit in law of the sequence $n\mathrm{tr}(Q\widehat{P}_n)$ and then we solve the optimization problem. Since

$$\begin{aligned} n\mathrm{tr}(Q\widehat{P}_n) &= n\mathrm{tr}(\widehat{\eta}^T Q\widehat{\eta}\ (\widehat{\eta}^T\widehat{\eta})^{-1}) \\ &= \mathrm{tr}(\sqrt{n}(\widehat{\eta}^T - \eta^T) Q \sqrt{n}(\widehat{\eta} - \eta)(\widehat{\eta}^T\widehat{\eta})^{-1}), \end{aligned}$$

Slutsky's theorem and the continuity of the operator $\mathrm{tr}(\cdot)$ provides that $n\mathrm{tr}(Q\widehat{P}_n)$ converges to $\mathrm{tr}(\delta^T Q \delta)$ 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. Thus it remains to calculate the expectation of this limit, notice that

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

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 $\mathrm{var}(Z\psi_k(Y))$ and using the linearity condition, we have

$$\mathbb{E}\left[W_\alpha\right] = \sum_{k=1}^{d} \mathbb{E}\left[\|QZ\|^2 \psi_k(Y)^2\right]. \tag{8}$$

Now let us reformulate the minimization problem in terms of matrix α. First, from (8) and using that the $I(h)$ are pairwise disjoint, we have

$$\mathbb{E}\left[W_\alpha\right] = \sum_{k=1}^{d} \alpha_k^T \mathbb{E}[\|Q_c Z\|^2 \mathbb{1}_Y \mathbb{1}_Y^T] \alpha_k = \mathrm{tr}(\alpha^T D \alpha), \tag{9}$$

and also,

$$\eta^T \eta = \alpha^T C^T C \alpha = (D^{\frac{1}{2}}\alpha)^T G D^{\frac{1}{2}} \alpha. \tag{10}$$

From (9) and (10) we set out the equivalent minimization problem

$$\min_{\alpha}\ \mathrm{tr}\left(\alpha^T D \alpha\right) \quad \text{u.c.} \quad (D^{\frac{1}{2}}\alpha)^T G D^{\frac{1}{2}} \alpha = Id,$$

then, from the variable change $U = P^T D^{\frac{1}{2}} \alpha$ we derive

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

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

$$(11)\quad \min_{U_1} \quad \mathrm{tr}(U_1 U_1^T) \quad \text{u.c.} \quad U_1 U_1^T = D_0,$$

where $U_1 \in \mathbb{R}^{d\times d}$. Clearly, in (11) the quantity to minimize is fixed by the constraint. Then, a solution of it is given by $U_1 = D_0^{-\frac{1}{2}} H$ where H is any orthogonal matrix. Hence, the solution of (7) is

$$(12)\quad \alpha = D^{-\frac{1}{2}} PU = D^{-\frac{1}{2}} P_1 D_0^{-\frac{1}{2}} H$$

where H is any orthogonal matrix. □

Proposition 1 provides the expression of the optimal functions $\psi_1, ..., \psi_d$. It is easy to show that their associated vectors $\eta = (\eta_1, ..., \eta_d)$ are such that

$$M_{TF1}\eta = \eta D_0,$$

where $M_{TF1} = \sum_{h=1}^{H} \frac{C_h C_h^T}{d_h}$. Hence we propose to follow this algorithm :

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

$$\widehat{d}_h = \frac{1}{n}\sum_{i=1}^{n} \|\widehat{Q}_c Z_i\|^2 \mathbb{1}_{\{Y_i \in I(h)\}}, \quad \widehat{C}_h = \frac{1}{n}\sum_{i=1}^{n} Z_i \mathbb{1}_{\{Y_i \in I(h)\}} \quad \text{and} \quad \widehat{M}_{TF1} = \sum_{h=1}^{H} \frac{\widehat{C}_h \widehat{C}_h^T}{\widehat{d}_h}.$$

2. Extract $\widehat{\eta} = (\widehat{\eta}_1, ..., \widehat{\eta}_d)$: the d eigenvectors of $\widehat{M}_{TF1}$ with largest eigenvalues.
3. $\widehat{Q}_c = I - \widehat{\eta}\widehat{\eta}^T$.

Steps 1 to 3 are repeated until convergence is achieved and then $\widehat{\eta}$ is the estimated basis of E_c derived from TF1. Simulations comparing TF1 to SIR will be presented at the oral presentation.

REFERENCES

[1] Efstathia Bura, *Dimension reduction via parametric inverse regression*, L_1-statistical procedures and related topics (Neuchatel, 1997), IMS Lecture Notes Monogr. Ser., vol. 31, Inst. Math. Statist., Hayward, CA, 1997, pp. 215–228. MR 1833590

[2] R. Dennis Cook, *Regression graphics*, Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons Inc., New York, 1998, Ideas for studying regressions through graphics, A Wiley-Interscience Publication. MR 1645673 (99i:62001)

[3] R. Dennis Cook and Liqiang Ni, *Sufficient dimension reduction via inverse regression: a minimum discrepancy approach*, J. Amer. Statist. Assoc. **100** (2005), no. 470, 410–428. MR MR2160547

[4] Y. Coudène, *Une version mesurable du théorème de Stone-Weierstrass*, Gaz. Math. (2002), no. 91, 10–17. MR 1896063 (2003c:28004)

[5] Ker-Chau Li, *Sliced inverse regression for dimension reduction*, J. Amer. Statist. Assoc. **86** (1991), no. 414, 316–342, With discussion and a rejoinder by the author. MR MR1137117 (93f:62084)

[6] Li-Xing Zhu and Kai-Tai Fang, *Asymptotics for kernel estimate of sliced inverse regression*, Ann. Statist. **24** (1996), no. 3, 1053–1068. MR MR1401836 (97k:60073)