HAL open science

A Note on Sliced Inverse Regression with Regularizations

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard

To cite this version:

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard. A Note on Sliced Inverse Regression with Regularizations. Biometrics, 2008, 64 (3), pp.982-986. 10.1111/j.1541-0420.2008.01080.x . inria-00180496v2

HAL Id: inria-00180496

https://inria.hal.science/inria-00180496v2

Submitted on 23 Apr 2013

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A Note on Sliced Inverse Regression with Regularizations

Caroline Bernard-Michel, Laurent Gardes, Stéphane Girard*

Laboratoire Jean-Kuntzmann & INRIA Rhône-Alpes, team Mistis,
Inovallée, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France,
(* corresponding author, Stephane.Girard@inrialpes.fr)

Abstract

In "Li, L. and Yin, X. (2008). Sliced Inverse Regression with Regularizations. *Biometrics*, 64(1):124–131" a ridge SIR estimator is introduced as the solution of a minimization problem and computed thanks to an alternating least-squares algorithm. This methodology reveals good performance in practice. In this note, we focus on the theoretical properties of the estimator. Is it shown that the minimization problem is degenerated in the sense that only two situations can occur: Either the ridge SIR estimator does not exist or it is zero.

Keywords: Inverse regression, regularization, sufficient dimension reduction.

1 Introduction

Many methods have been developed for inferring the conditional distribution of an univariate response Y given a predictor X in $\mathbb{R}^p$. When p is large, sufficient dimension reduction aims at replacing the predictor X by its projection onto a subspace of smaller dimension without loss of information on the conditional distribution of Y given X. In this context, the central subspace, denoted by $\mathcal{S}_{Y|X}$ plays an important role. It is defined as the smallest subspace such that, conditionally on the projection of X on $\mathcal{S}_{Y|X}$, Y and X are independent. In other words, the projection of X on $\mathcal{S}_{Y|X}$ contains all the information on Y that is available in the predictor X. Introducing $d = \dim(\mathcal{S}_{Y|X})$ and $A \in \mathbb{R}^{p\times d}$ such that $\mathcal{S}_{Y|X} = \mathrm{Span}(A)$, this property can be rewritten in terms of conditional distribution functions as

$$F(Y|X) = F(Y|A^T X).$$

The estimation of A has received considerable attention, and among the proposed methods, Sliced Inverse Regression (SIR) [6] seems to be the most popular one. Let us recall its definition from the minimum discrepancy point of view [3, 4]. Starting from a n- sample, and denoting by $\bar{X}$ the average of X, $\hat{\Sigma}_x$ the sample covariance matrix of X and assuming that the response variable Y is partitioned into h non-overlapping slices, the SIR estimator of A is obtained by minimizing

$$G(A, C) = \sum_{y=1}^{h} f_y \left((\bar{X}_y - \bar{X}) - \hat{\Sigma}_x A C_y \right)^T \hat{\Sigma}_x^{-1} \left((\bar{X}_y - \bar{X}) - \hat{\Sigma}_x A C_y \right) \qquad (1)$$

where $f_y = n_y/n$, n_y is the number of observations in the yth slice, $\bar{X}_y$ is the average of X in the yth slice and $C = (C_1, \ldots, C_h) \in \mathbb{R}^{d\times h}$. Defining

$$\hat{\Gamma} = \sum_{y=1}^{h} f_y (\bar{X}_y - \bar{X})(\bar{X}_y - \bar{X})^T,$$

an estimator of $\mathrm{cov}(\mathbb{E}(X|Y))$, the SIR estimator is obtained by computing the eigenvectors of $\hat{\Sigma}_x^{-1}\hat{\Gamma}$ associated to the d largest eigenvalues. It thus requires the inversion of $\hat{\Sigma}_x$ which is not possible as soon as $p > n$ or when the predictors are highly correlated. In order to overcome this problem, it has been proposed to use the ridge SIR estimator ([7], Definition 1) defined as follows. Let $\tau \geq 0$ and

$$G_\tau(A, C) = \sum_{y=1}^{h} f_y \|(\bar{X}_y - \bar{X}) - \hat{\Sigma}_x A C_y\|^2 + \tau \|\mathrm{vec}(A)\|^2, \tag{2}$$

where vec(.) is a matrix operator that stacks all columns of the matrix to a single vector. The ridge SIR estimator of the central subspace $\mathcal{S}_{Y|X}$ is $\mathrm{Span}(\hat{A})$ where

$$(\hat{A}, \hat{C}) = \arg\min_{A,C} G_\tau(A, C). \tag{3}$$

From the practical point of view, an alternating least-squares algorithm is proposed to solve this optimization problem [7]. It revealed good performances on simulated and real data. Here, we focus on the theoretical aspects. To this end, let us highlight that definition (3) assumes the existence of a unique minimum of G_τ. In Section 2, we prove that this is not the case. In fact, either $\arg\min G_\tau = \emptyset$, and thus the ridge SIR estimator does not exist, or $\arg\min G_\tau \subset \{0\} \times \mathbb{R}^{d\times h}$ and consequently the ridge SIR estimator is zero. A modification of the criterion (2) is proposed in Section 3 leading to the estimator of A proposed in [9]. Proofs are postponed to the Appendix.

2 On the existence of the ridge SIR estimator

Before stating our main result on the existence of the ridge SIR estimator, remark that G_τ ($\tau > 0$) does not penalize the same way two proportional matrices A and λA, $\lambda \in \mathbb{R} \setminus \{0\}$, although defining the same central subspace since $\mathrm{Span}(A) = \mathrm{Span}(\lambda A)$. This lack of invariance may explain why the ridge

SIR estimator is ill-defined as illustrated below.

Proposition 1. *Let $\tau > 0$. If $\arg\min G_\tau \neq \emptyset$ then $\hat{A}$ defined by (3) is the zero $p \times d$ matrix. Moreover,*

$$G_\tau(\hat{A}, C) = G_\tau(0, C) = \sum_{y=1}^{h} f_y \|\bar{X}_y - \bar{X}\|^2, \tag{4}$$

for all $C \in \mathbb{R}^{d\times h}$,

Since (4) does not depend on C, it follows that either $\arg\min G_\tau = \emptyset$ or $\arg\min G_\tau \subset \{0\} \times \mathbb{R}^{d\times h}$. The following proposition permits to distinguish between the two cases.

Proposition 2. *Let $\tau > 0$ and assume $\operatorname{rank}(\hat{\Sigma}_x) \geq d$. Then, $\arg\min G_\tau = \emptyset$ if and only if there exists $y \in \{1, \ldots, h\}$ such that $\hat{\Sigma}_x(\bar{X}_y - \bar{X}) \neq 0$.*

To solve the optimization problem (2), Li and Yin [7] proposed an alternating least-squares algorithm. At iteration $k+1$, given $A^{(k)}$, $C^{(k+1)}$ and $A^{(k+1)}$ are updated as:

$$\begin{aligned}
C_y^{(k+1)} &= \left(A^{(k)T}\hat{\Sigma}_x^2 A^{(k)}\right)^{-1} A^{(k)T}\hat{\Sigma}_x(\bar{X}_y - \bar{X}),\ y = 1, \ldots, h, \\
\operatorname{vec}\left(A^{(k+1)}\right) &= \left\{\sum_{y=1}^{h} f_y \left(C_y^{(k+1)T} \otimes \hat{\Sigma}_x\right)^T \left(C_y^{(k+1)T} \otimes \hat{\Sigma}_x\right) + \tau I_{pd}\right\}^{-1} \\
&\times \sum_{y=1}^{h} f_y \left(C_y^{(k+1)T} \otimes \hat{\Sigma}_x\right)^T (\bar{X}_y - \bar{X}).
\end{aligned}$$

The authors claimed that such an algorithm converges. As a consequence of Proposition 1, it is easily seen that the limit is always degenerated.

Corollary 1. *Let $\tau > 0$ and denote by (A^*, C^*) the limit of the sequence $\left(A^{(k)}, C^{(k)}\right)_k$. Necessarily, A^* is the zero $p \times d$ matrix.*

In view of this result, the good behavior of this algorithm on simulated and real data reported in [7], Section 3 and Section 4 cannot be justified from a theoretical point of view.

3 An alternative ridge SIR estimator

It is possible to modify the criterion G as follows

$$H_\tau(A,C) = G(A,C) + \tau \sum_{y=1}^{h} f_y \|AC_y\|^2. \tag{5}$$

The first advantage of H_τ is to be invariant with respect to bijective transformations, *i.e.*

$$H_\tau(AM, M^{-1}C) = H_\tau(A,C),$$

for all regular $d \times d$ matrix M. This property is natural since $\mathrm{span}(MA) = \mathrm{span}(A)$. Second, it is readily seen that the minimization of H_τ does not require the existence of $\hat{\Sigma}_x^{-1}$ since H_τ can be rewritten as

$$H_\tau(A,C) - H_\tau(0,0) = \sum_{y=1}^{h} f_y C_y^T A^T (\hat{\Sigma}_x + \tau I_p) AC_y - 2\sum_{y=1}^{h} f_y (\bar{X}_y - \bar{X})^T AC_y.$$

Finally, remarking that the original criterion G of SIR (1) can also be expanded as

$$G(A,C) - G(0,0) = \sum_{y=1}^{h} f_y C_y^T A^T \hat{\Sigma}_x AC_y - 2\sum_{y=1}^{h} f_y (\bar{X}_y - \bar{X})^T AC_y,$$

it appears that $H_\tau(A,C) - H_\tau(0,0)$ can be deduced from $G(A,C) - G(0,0)$ by substituting $\hat{\Sigma}_x + \tau I_p$ to $\hat{\Sigma}_x$. Consequently, the estimator of A obtained by minimizing (5) is the Regularized SIR estimator introduced in [9], generalized in [2] and applied in [1] since its columns are the eigenvectors of $(\hat{\Sigma}_x + \tau I_p)^{-1}\hat{\Gamma}$ associated to the d largest eigenvalues. As a conclusion, the introduction of the new functional (5) provides a theoretical framework for the Regularized SIR estimator [9]. Thus, a crossvalidation criterion could be derived, similarly to (8) in [7], for selecting the regularization parameter τ.

References

[1] Bernard-Michel, C., Fauvel, M., Gardes, L. and Girard, S. (2009). Retrieval of Mars surface physical properties from OMEGA hyperspectral images using Regularized Sliced Inverse Regression. *Journal of Geophysical Research - Planets*, **114**, E06005.

[2] Bernard-Michel, C., Gardes, L. and Girard, S. (2009). Gaussian Regularized Sliced Inverse Regression. *Statistics and Computing*, **19**, 85–98.

[3] Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Annals of Statistics*, **32**, 1061–1092.

[4] Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, **100**, 410–428.

[5] Harville, D. A. (1997). *Matrix algebra from a statistician's perspective*, Springer, New York.

[6] Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–327.

[7] Li, L. and Yin, X. (2008). Sliced Inverse Regression with Regularizations. *Biometrics*, **64**(1), 124–131.

[8] Magnus, J. R. and Neudecker, H. (1988). *Matrix differential calculus with applications in Statistics and Econometrics*, Wiley, New York.

[9] Zhong, W., Zeng, P., Ma, P., Liu, J. S. and Zhu, Y. (2005). RSIR: regularized sliced inverse regression for motif discovery. *Bioinformatics*, **21**(22), 4169–4175.

Appendix

Proof of Proposition 1 – Let us remark that

$$\begin{aligned} G_\tau(A,C) &= \sum_{y=1}^{h} f_y \left(\|\bar{X}_y - \bar{X}\|^2 - 2(\bar{X}_y - \bar{X})^T \hat{\Sigma}_x A C_y + C_y^T A^T \hat{\Sigma}_x^2 A C_y \right) \\ &+ \tau \|\mathrm{vec}(A)\|^2. \end{aligned} \tag{6}$$

Using the equality (see for instance [5], Chapter 16, equation (2.13)),

$$\hat{\Sigma}_x A C_y = (C_y^T \otimes \hat{\Sigma}_x)\mathrm{vec}(A), \tag{7}$$

for all $y = 1, \ldots, h$ and denoting $\tilde{a} = \mathrm{vec}(A)$, we thus have:

$$\begin{aligned} G_\tau(A,C) = G_\tau^*(\tilde{a},C) &= \sum_{y=1}^{h} f_y \Big\{ \|\bar{X}_y - \bar{X}\|^2 - 2(\bar{X}_y - \bar{X})^T (C_y^T \otimes \hat{\Sigma}_x)\tilde{a} \\ &+ \tilde{a}^T (C_y^T \otimes \hat{\Sigma}_x)^T (C_y^T \otimes \hat{\Sigma}_x)\tilde{a} \Big\} + \tau \|\tilde{a}\|^2. \end{aligned}$$

Suppose $\arg\min G_\tau \neq \emptyset$ and consider

$$(\hat{A}, \hat{C}) \in \arg\min_{A,C} G_\tau(A,C).$$

From [8], pp. 119-120, it follows that, necessarily, $(\hat{A}, \hat{C})$ is a stationary point of G_τ and thus satisfy the set of equations:

$$\nabla_i G_\tau^*(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) = 0, \ i = 1, \ldots, h+1, \tag{8}$$

where $\hat{a} = \mathrm{Vec}(\hat{A})$, $\hat{C} = (\hat{C}_1, \ldots, \hat{C}_h)$ and ∇_i denotes the gradient of G_τ^* with respect to its ith argument, $i = 1, \ldots, h+1$. Straightforward calculations lead to:

$$\begin{aligned} \nabla_1 G_\tau^*(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) &= 2 \sum_{y=1}^{h} f_y \Big\{ (\hat{C}_y^T \otimes \hat{\Sigma}_x)^T (\hat{C}_y^T \otimes \hat{\Sigma}_x)\hat{a} \\ &- (\hat{C}_y^T \otimes \hat{\Sigma}_x)^T (\bar{X}_y - \bar{X}) \Big\} + 2\tau \hat{a}, \end{aligned} \tag{9}$$

and, for $y = 1, \ldots, h$,

$$\begin{aligned}\nabla_{y+1} G^*_\tau(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) &= \nabla_{y+1} G_\tau(\hat{A}, \hat{C}_1, \ldots, \hat{C}_h) \\ &= 2 f_y \left\{ \hat{A}^T \hat{\Sigma}_x^2 \hat{A} \hat{C}_y - \hat{A}^T \hat{\Sigma}_x (\bar{X}_y - \bar{X}) \right\}. \quad (10)\end{aligned}$$

Thus, multiplying of the left by $\hat{C}_y^T$ and using (7), it follows

$$\begin{aligned}\hat{C}_y^T \nabla_{y+1} G^*_\tau(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) &= 2 f_y \left\{ \hat{C}_y^T (\hat{A}^T \hat{\Sigma}_x^2 \hat{A}) \hat{C}_y - \hat{C}_y^T \hat{A}^T \hat{\Sigma}_x (\bar{X}_y - \bar{X}) \right\} \\ &= 2 f_y \left\{ \hat{a}^T (\hat{C}_y^T \otimes \hat{\Sigma}_x)^T (\hat{C}_y^T \otimes \hat{\Sigma}_x) \hat{a} \right. \\ &- \left. \hat{a}^T (\hat{C}_y^T \otimes \hat{\Sigma}_x)^T (\bar{X}_y - \bar{X}) \right\}. \quad (11)\end{aligned}$$

Hence, collecting (9) and (11), it appears that

$$\hat{a}^T \nabla_1 G^*_\tau(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) = \sum_{y=1}^{h} \hat{C}_y^T \nabla_{y+1} G^*_\tau(\hat{a}, \hat{C}_1, \ldots, \hat{C}_h) + 2\tau \|\hat{a}\|^2.$$

Since the regularization parameter τ is positive, condition (8) implies $\|\hat{a}\|^2 = 0$, *i.e.* $\hat{A}$ is the zero $p \times d$ matrix. Replacing in (6), we have

$$G_\tau(\hat{A}, C) = G_\tau(0, C) = \sum_{y=1}^{h} f_y \|\bar{X}_y - \bar{X}\|^2,$$

for all $C \in \mathbb{R}^{d \times h}$ and the result is proved. ■

Proof of Corollary 1 – The limit of the sequence verifies the set of equations

$$\begin{aligned}C^*_y &= (A^{*T} \hat{\Sigma}_x^2 A^*)^{-1} A^{*T} \hat{\Sigma}_x (\bar{X}_y - \bar{X}), \; y = 1, \ldots, h, \\ \mathrm{vec}\,(A^*) &= \left\{ \sum_{y=1}^{h} f_y \left(C_y^{*T} \otimes \hat{\Sigma}_x \right)^T \left(C_y^{*T} \otimes \hat{\Sigma}_x \right) + \tau I_{pd} \right\}^{-1} \\ &\times \sum_{y=1}^{h} f_y \left(C_y^{*T} \otimes \hat{\Sigma}_x \right)^T (\bar{X}_y - \bar{X}).\end{aligned}$$

Thus, from (10) it follows that

$$\nabla_{y+1} G_\tau(A^*, C^*_1, \ldots, C^*_h) = 0$$

for all $y = 1, \ldots, h$, while, from (9),

$$\nabla_1 G_\tau(A^*, C_1^*, \ldots, C_h^*) = 0.$$

Consequently, (A^*, C^*) is a stationary point of G_τ, and, following the proof of Proposition 1, necessarily A^* is the zero $p \times d$ matrix. ■

Proof of Proposition 2 – First, let us suppose that $\hat{\Sigma}_x(\bar{X}_y - \bar{X}) = 0$ for all $y \in \{1, \ldots, h\}$. Then,

$$\begin{aligned} G_\tau(A, C) &= \sum_{y=1}^{h} f_y \|\bar{X}_y - \bar{X}\|^2 + \sum_{y=1}^{h} f_y C_y^T (A^T \hat{\Sigma}_x^2 A) C_y + \tau \|\mathrm{vec}(A)\|^2 \\ &\geq \sum_{y=1}^{h} f_y \|\bar{X}_y - \bar{X}\|^2 \\ &= G_\tau(0, C), \end{aligned}$$

which entails that $G_\tau(A, C)$ is minimum for every C if A is the zero matrix. As a consequence $\arg\min G_\tau \neq \emptyset$. This concludes the first part of the proof. Conversely, suppose there exists $y_0 \in \{1, \ldots, h\}$ such that $\hat{\Sigma}_x(\bar{X}_{y_0} - \bar{X}) \neq 0$. Let $\tau > 0$ and let us prove that there exist $A \in \mathbb{R}^{p \times d}$ and $C \in \mathbb{R}^{d \times h}$ such that $G_\tau(A, C) < G_\tau(0, C)$. To this end, let q_i, $i = 1, \ldots, p$ be the eigenvectors of $\hat{\Sigma}_x$ associated to the eigenvalues λ_i, $i = 1, \ldots, p$. Since

$$\hat{\Sigma}_x(\bar{X}_{y_0} - \bar{X}) = \sum_{i=1}^{p} \lambda_i q_i q_i^T (\bar{X}_{y_0} - \bar{X}) \neq 0,$$

there exists an eigenvector q^* associated to a random value $\lambda^* > 0$ such that $q^* {q^*}^T (\bar{X}_{y_0} - \bar{X}) \neq 0$. Thus $\|(\bar{X}_{y_0} - \bar{X})^T q^*\| \neq 0$, and let ε such that:

$$0 < \varepsilon < \sqrt{\frac{f_{y_0}}{\tau d}} \|(\bar{X}_{y_0} - \bar{X})^T q^*\|. \tag{12}$$

The matrices A and C are defined as follows. The first column of A is the vector εq^* and the $d-1$ following columns of A are the vectors εq_{j_i}, $i = 1, \ldots, d-1$ where the q_{j_i}'s are orthogonal eigenvectors (with unit norm) of $\hat{\Sigma}_x$ associated

to positive eigenvalues λ_{j_i}'s. Note that, since $\mathrm{rank}(\hat{\Sigma}_x) \geq d$, such a matrix A always exists. All the columns of C are chosen to be the null vector except the y_0th one defined by:

$$C_{y_0} = (A^T\hat{\Sigma}_x^2 A)^{-1} A^T \hat{\Sigma}_x (\bar{X}_{y_0} - \bar{X}) = \frac{1}{\varepsilon}\left(\frac{q^*}{\lambda^*}, \frac{q_{j_1}}{\lambda_{j_1}}, \ldots, \frac{q_{j_{d-1}}}{\lambda_{j_{d-1}}}\right)^T (\bar{X}_{y_0} - \bar{X}).$$

Such choices entail

$$\begin{aligned}
& G_\tau(A, C) - G_\tau(0, C) \\
= \; & \sum_{y=1}^{h} f_y \left\{ C_y^T (A^T\hat{\Sigma}_x^2 A) C_y - 2(\bar{X}_y - \bar{X})^T \hat{\Sigma}_x A C_y \right\} + \tau \|\mathrm{vec}(A)\|^2 \\
= \; & f_{y_0} \left\{ C_{y_0}^T (A^T\hat{\Sigma}_x^2 A) C_{y_0} - 2(\bar{X}_{y_0} - \bar{X})^T \hat{\Sigma}_x A C_{y_0} \right\} + \tau \|\mathrm{vec}(A)\|^2 \\
= \; & -f_{y_0} (\bar{X}_{y_0} - \bar{X})^T \hat{\Sigma}_x A (A^T\hat{\Sigma}_x^2 A)^{-1} A^T \hat{\Sigma}_x (\bar{X}_{y_0} - \bar{X}) + \tau \|\mathrm{vec}(A)\|^2 \\
= \; & -f_{y_0} \|(\bar{X}_{y_0} - \bar{X})^T q^*\|^2 - f_{y_0} \sum_{i=1}^{d-1} \|(\bar{X}_{y_0} - \bar{X})^T q_{j_i}\|^2 + \tau d \varepsilon^2 \\
\leq \; & -f_{y_0} \|(\bar{X}_{y_0} - \bar{X})^T q^*\|^2 + \tau d \varepsilon^2 \\
< \; & 0,
\end{aligned}$$

from (12). Thus, $(0, C) \notin \arg\min G_\tau$ and taking account of Proposition 1 yields $\arg\min G_\tau = \emptyset$. ■