Journal of Statistical Planning and Inference 139 (2009) 3570 – 3578

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

The sliced inverse regression algorithm as a maximum likelihood procedure☆

María Eugenia Szretter[a], Víctor Jaime Yohai[b,*]

[a] Universidad de Buenos Aires, Argentina
[b] Departamento de Matemáticas, Universidad de Buenos Aires, Pabellón I—Ciudad Universitaria, 1428 Buenos Aires, and CONICET, Argentina

ARTICLE INFO

Article history:
Received 6 December 2008
Received in revised form
5 April 2009
Accepted 7 April 2009
Available online 18 April 2009

Keywords:
Sliced inverse regression
Maximum likelihood estimates
Central mean subspace

ABSTRACT

It is shown that the sliced inverse regression procedure proposed by Li corresponds to the maximum likelihood estimate where the observations in each slice are samples of multivariate normal distributions with means in an affine manifold.

1. Introduction

Regression models are used to describe a relationship between a response variable y and a group of covariates $\mathbf{x}=(x_1,\ldots,x_p)\in\mathbb{R}^p$. When there is not an adequate parametric model that gives a satisfactory fit to the data, non-parametric techniques appear as more flexible tools. However, when p increases the number of observations required to use local smoothing techniques increases exponentially, and consequently these methods become unfeasible. This limitation of the non-parametric techniques is known in the statistical literature as the *dimensional curse*.

One way to overcome this difficulty is to use models where y depends in a non-parametric way only on a few projections of the vector $\mathbf{x}$ along some directions, i.e., the response y depends on $\mathbf{x}$ only through $\boldsymbol{\beta}_1'\mathbf{x},\ldots,\boldsymbol{\beta}_K'\mathbf{x}$, where K is much smaller than p and $\boldsymbol{\beta}_i$, $1\leqslant i\leqslant K$ are vectors on $\mathbb{R}^p$ that can without loss of generality be assumed to have norm one.

Li (1991) considers the following model:

$$y=g(\boldsymbol{\beta}_1'\mathbf{x},\ldots,\boldsymbol{\beta}_K'\mathbf{x},\varepsilon), \tag{1}$$

where g is a smooth function and the error ε is independent of $\mathbf{x}$. Note that once the directions $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K$ are known the local smoothing procedure to determine g involves only K variables, and therefore if K is small the dimensionality curse is overcome.

A more general approach to the problem of dimension reduction can be found in Cook and Li (2002). Let $\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_K$ vectors in $\mathbb{R}^p$, then the subspace $\mathscr{V}$ generated by these vectors is called a *mean dimension-reduction subspace* if

$$E(y|\mathbf{x})=g(\boldsymbol{\beta}_1'\mathbf{x},\ldots,\boldsymbol{\beta}_K'\mathbf{x}). \tag{2}$$

☆ This research was partially supported by Grants X-094 from Universidad de Buenos Aires, PID 5505 from CONICET and PAV 120 and PICT 21407 from ANPCYT, Argentina.

* Corresponding author. Tel.: +54 11 4576 3375.
E-mail addresses: meszre@dm.uba.ar (M.E. Szretter), vyohai@uolsinectis.com.ar, vyohai@mate.dm.uba.ar (V.J. Yohai).

0378-3758/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2009.04.008

Note that the validity of (2) does not depend on the base of $\mathscr{V}$ which is used. Let

$$\mathscr{V}_0 = \cap \mathscr{V}_m,$$

where the intersection is overall mean dimension-reduction subspaces $\mathscr{V}_m$. If $\mathscr{V}_0$ is itself a mean dimension-reduction subspace, it is called *the central mean subspace* (CMS). Cook and Li (2002) gave mild conditions for the existence of $\mathscr{V}_0$. Additional existence results can be found in Cook (1994, 1996, 1998). Note that if the CMS exists, it should be unique. We will suppose in the remainder of this paper that $\mathscr{V}_0$ exists.

The CMS can be estimated by using the sliced inverse regression (SIR) algorithm proposed by Li (1991). The procedure consists of dividing the observations in slices according to the value of y belonging to intervals $I_1, \ldots, I_H$ and assuming that $E(\mathbf{x}|y \in I_h) - E(\mathbf{x})$, $1 \leqslant h \leqslant H$ belong to the space

$$\mathscr{V}_0^* = \{\boldsymbol{\gamma} = \Psi\boldsymbol{\beta} : \boldsymbol{\beta} \in \mathscr{V}_0\},$$

where Ψ is the covariance matrix of $\mathbf{x}$. We will also assume that we know the dimension K of $\mathscr{V}_0$. Several authors have studied how to determine K. Among them, we can cite Li (1991), Schott (1994) and Ferré (1998).

Since the seminal paper by Li (1991), several procedures using the inverse regression approach have been proposed. We can cite among many others: sliced average variance estimation (Cook and Weisberg, 1991), principal Hessian directions (Li, 1992; Cook, 1998), inverse regression (Cook and Ni, 2005), graphical methods (Cook, 1994; Cook and Wetzel, 1993), simple contour regression (Li et al., 2005), Fourier estimation (Zhu and Zeng, 2006) and principal fitted components (PFC) (Cook, 2007).

In this paper we show that the SIR procedure is equivalent to estimate a K-dimensional subspace $\mathscr{V}_0 \subset \mathbb{R}^p$ by maximum likelihood assuming that the distribution of $\mathbf{x}$ given that $y \in I_h$ is $\mathrm{N}_p(\boldsymbol{\mu}_h, \Sigma)$, where $\mathrm{N}_p(\boldsymbol{\mu}, \Sigma)$ denotes the p-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ, and that the conditional means $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H$ belong to an affine manifold of dimension K. In Section 3, we discuss the connection between this result and those in Cook (2007).

In Section 2 we describe the SIR algorithm. In Section 3 we state the main results regarding the equivalence between the estimates of the subspace obtained with the SIR algorithm and the maximum likelihood estimates. The Appendix contains proofs.

2. Sliced inverse regression model

Let $(\mathbf{x}_i, y_i)$ with $1 \leqslant i \leqslant N$ be a random sample of a distribution satisfying (2). Call Ψ the covariance matrix of $\mathbf{x}$. Let $\overline{\mathbf{x}}_{\cdot\cdot}$ and $\widehat{\Psi}$ be the sample mean and sample covariance estimates of the $\mathbf{x}_i$'s, respectively. The SIR algorithm is as follows:

(1) Standardize $\mathbf{x}_i$ by computing $\mathbf{z}_i = \widehat{\Psi}^{-1/2}[\mathbf{x}_i - \overline{\mathbf{x}}_{\cdot\cdot}]$.
(2) Divide the range of y in H slices $I_1, \ldots, I_H$, where $I_h = (\zeta_h, \zeta_{h+1}]$, $-\infty = \zeta_1 < \zeta_2 < \cdots < \zeta_{H+1} = \infty$. Let $(\mathbf{z}_{hj}, y_{hj})$, $1 \leqslant j \leqslant n_h$, be the standardized observations of the sample whose y value falls in I_h.
(3) For each slice, compute $\overline{\mathbf{z}}_h$, the sample mean of the observations $\mathbf{z}_i$ whose y value falls in I_h and perform a (weighted) principal components analysis for $\overline{\mathbf{z}}_h$ as follows. Compute the eigenvalues and eigenvectors of the weighted covariance matrix

$$\widehat{\Phi} = \frac{1}{N}\sum_{h=1}^{H} n_h \overline{\mathbf{z}}_h \overline{\mathbf{z}}_h'.$$

(4) Let $\widehat{\boldsymbol{\eta}}_1, \ldots, \widehat{\boldsymbol{\eta}}_K$ be the eigenvectors corresponding to the K largest eigenvalues of $\widehat{\Phi}$. The estimate of the CMS is the subspace spanned by

$$\widehat{\boldsymbol{\beta}}_k = \widehat{\Psi}^{-1/2}\widehat{\boldsymbol{\eta}}_k, \quad 1 \leqslant k \leqslant K. \tag{3}$$

3. The SIR estimate as a maximum likelihood procedure

We will show that the SIR procedure described in Section 2 is equivalent to estimate the CMS subspace by maximum likelihood assuming that the sliced samples $\mathbf{x}_{hj}$, $1 \leqslant j \leqslant n_h$, are independent random samples $\mathrm{N}_p(\boldsymbol{\mu}_h, \Sigma)$, where $\boldsymbol{\mu}_h$ belongs to an affine manifold of dimension K. Observe that Theorem 3.1 of Li (1991) implies that under some assumptions on the distribution of $\mathbf{x}$, the conditional means $\boldsymbol{\mu}_h = E(\mathbf{x}|y \in I_h)$ belongs to the affine manifold of dimension K, $\mathscr{V}_0^* + \mathbf{a}$, where $\mathbf{a} \in \mathbb{R}^p$ and can be taken equal to $E(\mathbf{x})$. To guarantee that the SIR algorithm estimates the CMS, we will assume the *coverage condition* that $\boldsymbol{\mu}_1 - E(\mathbf{x}), \ldots, \boldsymbol{\mu}_H - E(\mathbf{x})$ generate $\mathscr{V}_0^*$.

We need the following definitions:

$$\overline{\mathbf{x}}_{\cdot\cdot} = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h} \mathbf{x}_{hj},$$

$$\overline{\mathbf{x}}_{h\cdot} = \frac{1}{n_h}\sum_{j=1}^{n_h} \mathbf{x}_{hj},$$

$$B = \frac{1}{N}\sum_{h=1}^{H} n_h(\overline{\mathbf{x}}_{h\cdot} - \overline{\mathbf{x}}_{\cdot\cdot})(\overline{\mathbf{x}}_{h\cdot} - \overline{\mathbf{x}}_{\cdot\cdot})' \tag{4}$$

and

$$W = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{h\cdot})(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{h\cdot})'. \tag{5}$$

Note that $\overline{\mathbf{x}}_{\cdot\cdot}$ and $\overline{\mathbf{x}}_{h\cdot}$ are p-dimensional column vectors and B and W are $p \times p$ matrices.

Let C be the orthogonal matrix $[\mathbf{c}_1 \cdots \mathbf{c}_p]$, where $\mathbf{c}_i$ is an eigenvector of $B^{-1/2}WB^{-1/2}$ corresponding to the eigenvalue θ_i, and $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_p$. Denote by Θ the diagonal matrix with $\theta_1, \ldots, \theta_p$ in the diagonal, then

$$B^{-1/2}WB^{-1/2} = C\Theta C'.$$

Let C_h be the matrix with the first h columns of C.

Theorem 1. *Assume that for $1 \leqslant h \leqslant H$, $(\mathbf{x}_{hj})_{1 \leqslant j \leqslant n_h}$ are i.i.d. column vectors in $\mathbb{R}^p$ with distribution $\mathrm{N}_p(\boldsymbol{\mu}_h, \Sigma)$, where $\boldsymbol{\mu}_h$ belongs to $\mathscr{V}_0^* + \mathbf{a}$, $\mathscr{V}_0^*$ is a K-dimensional subspace of $\mathbb{R}^p$ and $\mathbf{a} \in \mathbb{R}^p$. We also assume that the H samples are independent. Then,*

(a) *The maximum likelihood estimate of Σ is*

$$\widehat{\Sigma} = W + B^{1/2}C_{p-K}C'_{p-K}B^{1/2}.$$

(b) *Let $\widehat{\mathbf{t}}_i$, $1 \leqslant i \leqslant p$, be orthogonal eigenvectors of norm one of $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ corresponding to the eigenvalues $\widehat{\lambda}_1 \geqslant \widehat{\lambda}_2 \geqslant \cdots \geqslant \widehat{\lambda}_p$. The maximum likelihood estimate of $\boldsymbol{\mu}_h$* is

$$\widehat{\boldsymbol{\mu}}_h = \widehat{\Sigma}^{1/2}\widehat{T}_K\widehat{T}'_K\widehat{\Sigma}^{-1/2}(\overline{\mathbf{x}}_{h\cdot} - \overline{\mathbf{x}}_{\cdot\cdot}) + \overline{\mathbf{x}}_{\cdot\cdot}, \quad 1 \leqslant h \leqslant H,$$

where $\widehat{T}_K = [\widehat{\mathbf{t}}_1 \cdots \widehat{\mathbf{t}}_K]$. Then $\widehat{\boldsymbol{\mu}}_h$ is the orthogonal projection, using the norm associated to $\widehat{\Sigma}$, of $\overline{\mathbf{x}}_{h\cdot} - \overline{\mathbf{x}}_{\cdot\cdot}$ on the K-dimensional affine manifold $\widehat{\mathscr{V}}^* + \overline{\mathbf{x}}_{\cdot\cdot}$, where $\widehat{\mathscr{V}}^*$ is the subspace spanned by $\widehat{\Sigma}^{1/2}\widehat{\mathbf{t}}_1, \ldots, \widehat{\Sigma}^{1/2}\widehat{\mathbf{t}}_K$.

(c) *$\widehat{\Sigma}^{1/2}\widehat{\mathbf{t}}_i$ is an eigenvector of BW^{-1} corresponding to the eigenvalue $1/\theta_{p-i+1}$.*

(d) *$\widehat{\Sigma}$ can also be written as*

$$\widehat{\Sigma} = \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'$$

and it satisfies

$$\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\widehat{\Sigma}^{-1}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h) = p. \tag{6}$$

Remark 1. We should mention that Cook (2007) proposed a general inverse regression model called principal fitted components (PFC) model. The PFC model assumes that the distribution of $\mathbf{x}$ given y is $\mathrm{N}_p(\boldsymbol{\mu}(y), \Sigma)$, where $\boldsymbol{\mu}(y) = \mathbf{a} + \Gamma\Lambda\mathbf{f}(y)$, where Γ and Λ are unknown matrices of dimension $p \times K$ and $K \times H$, respectively, $\mathbf{a} \in \mathbb{R}^p$ is an unknown vector and $\mathbf{f} : \mathbb{R} \to \mathbb{R}^H$ is a known function. Therefore $\boldsymbol{\mu}(y)$ belongs to the manifold $\mathscr{V} + \mathbf{a}$, where $\mathscr{V}$ is the subspace of dimension K generated by the columns of Γ. If the observations $\mathbf{x}_{hj}$, $1 \leqslant j \leqslant n_h$, are those for which $y \in I_h$ as in the SIR algorithm, the model assumed in Theorem 1 is the PFC model corresponding to $\mathbf{f}(y) = (1_{I_1}(y), \ldots, 1_{I_h}(y))'$, where 1_A denotes de indicator function of the set A. In Cook (2007) the maximum likelihood estimates of the manifold $\mathscr{V} + \mathbf{a}$ is obtained assuming that Σ is known. Instead, in Theorem 1 we derive the simultaneous maximum likelihood estimates of $\mathscr{V} + \mathbf{a}$ and Σ.

The following theorem establishes the relation between the maximum likelihood estimate of $\mathscr{V}_0^*$ given in Theorem 1 and the estimate of the CMS given by the SIR algorithm.

Theorem 2. *The estimate of the CMS obtained with the SIR algorithm coincides with $\widehat{\Psi}^{-1}\widehat{\mathscr{V}}^*$, where $\widehat{\mathscr{V}}^*$ is the maximum likelihood estimate of $\mathscr{V}_0^*$ given in* Theorem 1. *It also coincides with the subspace spanned by the K eigenvectors corresponding to the K largest eigenvalues of BW^{-1}.*

Acknowledgement

We thank to the referee for his valuable suggestions, which considerably improved the paper.

Appendix

Proof of Theorem 1. *Proof of* (b): Let $N=\sum_{h=1}^{H} n_h$. Then the logarithm of the likelihood is

$$\ln L(\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_H,\Sigma)=c-\frac{N}{2}\ln|\Sigma|-\frac{1}{2}\sum_{h=1}^{H}\sum\nolimits_{j=1}^{n_h}(\mathbf{x}_{hj}-\boldsymbol{\mu}_h)'\Sigma^{-1}(\mathbf{x}_{hj}-\boldsymbol{\mu}_h), \tag{7}$$

where $c=-(1/2)pN\ln 2\pi$. We start by estimating $\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_H$ assuming Σ known. Put $\boldsymbol{\mu}_h=\boldsymbol{\alpha}_h+\mathbf{a}$, where $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_H$ are in a K-dimensional subspace $\mathscr{V}^*\subset\mathbb{R}^P$ and $\mathbf{a}\in\mathbb{R}^P$. Let us find first $\mathbf{a}$ minimizing

$$\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\boldsymbol{\alpha}_h-\mathbf{a})'\Sigma^{-1}(\mathbf{x}_{hj}-\boldsymbol{\alpha}_h-\mathbf{a}) \tag{8}$$

for $\boldsymbol{\alpha}_h$, $1\leqslant h\leqslant H$ given. Put

$$\overline{\boldsymbol{\alpha}}=\frac{1}{N}\sum_{h=1}^{H}n_h\boldsymbol{\alpha}_h.$$

Then, it is easy to verify that (8) equals

$$\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\boldsymbol{\alpha}_h-(\overline{\mathbf{x}}_{\cdot\cdot}-\overline{\boldsymbol{\alpha}}))'\Sigma^{-1}(\mathbf{x}_{hj}-\boldsymbol{\alpha}_h-(\overline{\mathbf{x}}_{\cdot\cdot}-\overline{\boldsymbol{\alpha}}))+N(\mathbf{a}-(\overline{\mathbf{x}}_{\cdot\cdot}-\overline{\boldsymbol{\alpha}}))'\Sigma^{-1}(\mathbf{a}-(\overline{\mathbf{x}}_{\cdot\cdot}-\overline{\boldsymbol{\alpha}})),$$

and therefore $\mathbf{a}=\overline{\mathbf{x}}_{\cdot\cdot}-\overline{\boldsymbol{\alpha}}$. Since $\overline{\boldsymbol{\alpha}}\in\mathscr{V}^*$, and $\boldsymbol{\mu}_h=\boldsymbol{\alpha}_h-\overline{\boldsymbol{\alpha}}+\overline{\mathbf{x}}_{\cdot\cdot}$, redefining $\boldsymbol{\alpha}_h$ as $\boldsymbol{\alpha}_h-\overline{\boldsymbol{\alpha}}$ we have $\boldsymbol{\mu}_h=\boldsymbol{\alpha}_h+\overline{\mathbf{x}}_{\cdot\cdot}$, where the α_h's are in $\mathscr{V}^*$ too. Let $\boldsymbol{\alpha}_h^{\mathscr{V}^*}$, $1\leqslant h\leqslant H$, be the orthogonal projection of $\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}$ in $\mathscr{V}^*$ when $\mathbb{R}^p$ is endowed with the metric induced by the norm $\|\mathbf{x}\|_\Sigma^2=\mathbf{x}'\Sigma^{-1}\mathbf{x}$. Then, since the cross product terms are zero, we can write

$$\begin{aligned}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\boldsymbol{\mu}_h)'\Sigma^{-1}(\mathbf{x}_{hj}-\boldsymbol{\mu}_h)&=\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h)'\Sigma^{-1}\left(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h\right)\\&=\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{h\cdot})'\Sigma^{-1}(\mathbf{x}_{hj}-\overline{\mathbf{x}}_{h\cdot})+\sum_{h=1}^{H}n_h\left(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*}\right)'\Sigma^{-1}\left(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*}\right)\\&\quad+\sum_{h=1}^{H}n_h(\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_h^{\mathscr{V}^*})'\Sigma^{-1}\left(\boldsymbol{\alpha}_h-\boldsymbol{\alpha}_h^{\mathscr{V}^*}\right).\end{aligned} \tag{9}$$

Then, for fixed Σ, the minimum of (7) is obtained by taking

$$\boldsymbol{\mu}_h=\boldsymbol{\alpha}_h+\overline{\mathbf{x}}_{\cdot\cdot} \tag{10}$$

and $\mathscr{V}^*$ the K-dimensional subspace minimizing the second term of the right side of (9) given by

$$\sum_{h=1}^{H}n_h(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*})'\Sigma^{-1}\left(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*}\right). \tag{11}$$

Consider the transformation $\mathbf{v}_h=n_h^{1/2}\Sigma^{-1/2}(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot})$, $\mathscr{V}^{**}=\Sigma^{-1/2}\mathscr{V}^*$. Then $\gamma_h^{\mathscr{V}^{**}}=n_h^{1/2}\Sigma^{-1/2}\alpha_h^{\mathscr{V}^*}$ is the orthogonal projection of $\mathbf{v}_h$ in $\mathscr{V}^{**}$ with the metric induced by the identity matrix. Then finding $\mathscr{V}^*$ minimizing (11) is equivalent to finding $\mathscr{V}^{**}$ minimizing

$$\sum_{h=1}^{H}\left\|\mathbf{v}_h-\gamma_h^{\mathscr{V}^{**}}\right\|^2. \tag{12}$$

Then $\mathscr{V}^{**}$ is the solution to the principal components of the sample $\mathbf{v}_h$, $1\leqslant h\leqslant H$. Note that the sample covariance matrix of the $\mathbf{v}_h$'s is $(N/H)\Sigma^{-1/2}B\Sigma^{-1/2}$, where B is given in (4). Then, according to Theorem 5.5 of Seber (1986), (12) is minimized as follows. Let $\mathbf{t}_i$, $1\leqslant i\leqslant p$, be orthogonal eigenvectors of $\Sigma^{-1/2}B\Sigma^{-1/2}$ with norm one corresponding to the eigenvalues $\lambda_1\geqslant\lambda_2\geqslant\cdots\geqslant\lambda_p$ and let $T_K=[\mathbf{t}_1\cdots\mathbf{t}_K]$. Then $\mathscr{V}^{**}$ is the subspace spanned by $\mathbf{t}_i$, $1\leqslant i\leqslant K$, and

$$\begin{aligned}\min_{\mathscr{V}^{**}}\sum_{h=1}^{H}\left\|\mathbf{v}_h-\gamma_h^{\mathscr{V}^{**}}\right\|^2&=\min_{\mathscr{V}^{*}}\sum_{h=1}^{H}n_h(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*})'\Sigma^{-1}\left(\overline{\mathbf{x}}_{h\cdot}-\overline{\mathbf{x}}_{\cdot\cdot}-\boldsymbol{\alpha}_h^{\mathscr{V}^*}\right)\\&=N\sum_{i=K+1}^{p}\lambda_i.\end{aligned} \tag{13}$$

Besides

$$\gamma_h^{\mathscr{V}^*} = T_K T_K' \mathbf{v}_h.$$

Since $\mathscr{V}^* = \Sigma^{1/2}\mathscr{V}^{**}$ and $\boldsymbol{\alpha}_h^{\mathscr{V}^{**}} = n_h^{-1/2}\Sigma^{1/2}\gamma_h^{\mathscr{V}^*}$, we obtain that $\mathscr{V}^*$ is the subspace spanned by $\Sigma^{1/2}\mathbf{t}_i$, $1 \leqslant i \leqslant K$, and

$$\boldsymbol{\alpha}_h^{\mathscr{V}^*} = \Sigma^{1/2} T_K T_K' \Sigma^{-1/2}(\overline{\mathbf{x}}_{h\cdot} - \overline{\mathbf{x}}_{\cdot\cdot}).$$

This proves (b).

Proof of (a): Now we will find the maximum likelihood estimate of Σ. Given a $p \times p$ matrix A we denote by $\lambda_1(A) \geqslant \cdots \geqslant \lambda_p(A)$ the eigenvalues of A when they are all real. From (7), (9), (10) and (13) we have

$$\max_{\boldsymbol{\mu}_1 \in \mathscr{V}^* + \mathbf{a}, \ldots, \boldsymbol{\mu}_H \in \mathscr{V}^* + \mathbf{a}} \ln L(\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_H, \Sigma) = c - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{h\cdot})'\Sigma^{-1}(\mathbf{x}_{hj} - \overline{\mathbf{x}}_{h\cdot}) - \frac{N}{2}\sum_{i=K+1}^{p}\lambda_i(\Sigma^{-1}B)$$

$$= c - \frac{N}{2}\left[\ln|\Sigma| + \operatorname{trace}\{\Sigma^{-1}W\} + \sum_{i=K+1}^{p}\lambda_i(\Sigma^{-1}B)\right],$$

where W is given in (5). Then Σ should minimize

$$f(\Sigma) = \ln|\Sigma| + \operatorname{trace}(\Sigma^{-1}W) + \sum_{i=K+1}^{p}\lambda_i(\Sigma^{-1}B).$$

Put

$$\Sigma^* = B^{-1/2}\Sigma B^{-1/2} \tag{14}$$

and $A = B^{-1/2}WB^{-1/2}$. Then it is straightforward to show that

$$f(B^{1/2}\Sigma^* B^{1/2}) = \ln|B| + \ln|\Sigma^*| + \operatorname{trace}\{\Sigma^{*-1}(B^{-1/2}WB^{-1/2})\} + \sum_{i=K+1}^{p}\lambda_i(\Sigma^{*-1})$$

$$= \ln|B| + g(\Sigma^*),$$

where

$$g(\Sigma^*) = \ln|\Sigma^*| + \operatorname{trace}\{\Sigma^{*-1}A\} + \sum_{i=K+1}^{p}\lambda_i(\Sigma^{*-1}).$$

Then we have to find Σ^* minimizing $g(\Sigma^*)$.

We can write

$$\Sigma^* = D\Omega D' \tag{15}$$

with $\Omega = \operatorname{diag}(\omega_1, \ldots, \omega_p)$, $\omega_1 \geqslant \omega_2 \geqslant \cdots \geqslant \omega_p \geqslant 0$ the eigenvalues of Σ^* and $D = [\mathbf{d}_1 \cdots \mathbf{d}_p]$ the orthogonal matrix where $\mathbf{d}_i$ is an eigenvector corresponding to ω_i. Then

$$\ln|\Sigma^*| = \ln\left(\prod_{i=1}^{p}\omega_i\right) = \sum_{i=1}^{p}\ln(\omega_i),$$

$$\operatorname{trace}\{\Sigma^{*-1}A\} = \operatorname{trace}(D\Omega^{-1}D'A) = \operatorname{trace}(\Omega^{-1}D'AD)$$

$$= \sum_{i=1}^{p}\frac{\mathbf{d}_i'A\mathbf{d}_i}{\omega_i}$$

and

$$\sum_{i=K+1}^{p}\lambda_i(\Sigma^{*-1}) = \frac{1}{\omega_1} + \cdots + \frac{1}{\omega_{p-K}}.$$

Then we have to find the values of (D, Ω) minimizing

$$g^*(D,\Omega) = g(D\Omega D') = \sum_{i=1}^{p} \ln(\omega_i) + \sum_{i=1}^{p} \frac{\mathbf{d}_i' A \mathbf{d}_i}{\omega_i} + \sum_{i=1}^{p-K} \frac{1}{\omega_i}$$

subject to the constraints that D is an orthogonal matrix and Ω is a diagonal matrix, with positive diagonal elements.

Observe that the function $r(a) = \ln(a) + b/a$ is minimized by taking $a = b$. Then we have

$$\omega_i = \mathbf{d}_i' A \mathbf{d}_i + 1, \quad 1 \leqslant i \leqslant p - K, \tag{16}$$

$$\omega_i = \mathbf{d}_i' A \mathbf{d}_i, \quad p - K + 1 \leqslant i \leqslant p \tag{17}$$

and D is an orthogonal matrix minimizing

$$m(D) = \sum_{i=1}^{p-K} \ln(\mathbf{d}_i' A \mathbf{d}_i + 1) + \sum_{i=p-K+1}^{p} \ln(\mathbf{d}_i' A \mathbf{d}_i).$$

Consider the spectral decomposition $A = C\Theta C'$, where C is orthogonal and Θ diagonal with diagonal elements $\theta_1 \geqslant \theta_2 \geqslant \cdots \geqslant \theta_p$. Put

$$E = C'D, \tag{18}$$

then $E = [\mathbf{e}_1 \cdots \mathbf{e}_p]$ is an orthogonal matrix minimizing

$$m^*(E) = \sum_{i=1}^{p-K} \ln\left(\sum_{j=1}^{p} \theta_j e_{ji}^2 + 1\right) + \sum_{i=p-K+1}^{p} \ln\left(\sum_{j=1}^{p} \theta_j e_{ji}^2\right).$$

Since $\ln(x+1)$ and $\ln(x)$ are concave we have

$$\begin{aligned} m^*(E) &\geqslant \sum_{i=1}^{p-K} \sum_{j=1}^{p} e_{ji}^2 \ln(\theta_j + 1) + \sum_{i=p-K+1}^{p} \sum_{j=1}^{p} e_{ji}^2 \ln(\theta_j) \\ &= \sum_{j=1}^{p} f_j \ln(\theta_j + 1) + \sum_{j=1}^{p} (1 - f_j) \ln(\theta_j), \end{aligned} \tag{19}$$

where $f_j = \sum_{i=1}^{p-K} e_{ji}^2$. Clearly

$$0 \leqslant f_j \leqslant 1, \quad \sum_{j=1}^{p} f_j = p - K. \tag{20}$$

We are going to show that the minimum of

$$b(f_1, \ldots, f_p) = \sum_{j=1}^{p} f_j \ln(\theta_j + 1) + \sum_{j=1}^{p} (1 - f_j) \ln(\theta_j)$$

subject to (20) occurs when $f_j = 1$, for $1 \leqslant j \leqslant p - K$ and $f_j = 0$ for $p - K + 1 \leqslant j \leqslant p$. To prove this, it is enough to show that if $f_{i_0} < 1$ for some $1 \leqslant i_0 \leqslant p - K$ we can still decrease $b(f_1, \ldots, f_p)$. In fact in that case there exists $j_0 > p - K$ such that $f_{j_0} > 0$. Take $\varepsilon > 0$ such that $f_{i_0} + \varepsilon < 1$ and $f_{j_0} - \varepsilon > 0$ and put $f_{i_0}^* = f_{i_0} + \varepsilon, f_{j_0}^* = f_{j_0} - \varepsilon$ and $f_i^* = f_i$ for $i \neq i_0, j_0$.

Then

$$\begin{aligned} b(f_1, \ldots, f_p) - b(f_1^*, \ldots, f_p^*) &= -\varepsilon \ln(\theta_{i_0} + 1) + \varepsilon \ln(\theta_{i_0}) + \varepsilon \ln(\theta_{j_0} + 1) - \varepsilon \ln(\theta_{j_0}) \\ &= \varepsilon \ln\left(\frac{(\theta_{j_0} + 1)\theta_{i_0}}{(\theta_{i_0} + 1)\theta_{j_0}}\right) \\ &= \varepsilon \ln\left(\frac{1 + \dfrac{1}{\theta_{j_0}}}{1 + \dfrac{1}{\theta_{i_0}}}\right) \geqslant 0. \end{aligned}$$

Then by (19) we get

$$m^*(E) \geqslant \sum_{j=1}^{p-K} \ln(\theta_j + 1) + \sum_{j=p-K+1}^{p} \ln(\theta_j).$$

On the other hand

$$m^*(I_p) = \sum_{j=1}^{p-K} \ln(\theta_j + 1) + \sum_{j=p-K+1}^{p} \ln(\theta_j).$$

Then the identity matrix $E = I_p$ is an orthogonal matrix which minimizes (19) and therefore by (18) the orthogonal matrix $D = C$ minimizes $m(D)$. Then by (15)–(17) we get

$$\begin{aligned}\widehat{\Sigma}^* &= C\Theta C' + C_{p-K}C'_{p-K} \\ &= B^{-1/2}WB^{-1/2} + C_{p-K}C'_{p-K},\end{aligned}$$

where C_h denotes the matrix with the first h columns of C. Using (14) we get

$$\widehat{\Sigma} = W + B^{1/2}C_{p-K}C'_{p-K}B^{1/2},$$

and this proves (a).

Proof of (c): Take

$$\mathbf{z}_i = \widehat{\Sigma}^{-1/2}B^{1/2}\mathbf{c}_i,$$

where $\mathbf{c}_i$ is the eigenvector of $A = B^{-1/2}WB^{-1/2}$ corresponding to the eigenvalue θ_i. We start proving that $\mathbf{z}_i$ is an eigenvector of $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ corresponding to the eigenvalue $1/\gamma_i$ where γ_i is given by

$$\gamma_i = \begin{cases} \theta_i + 1 & \text{if} \quad 1 \leqslant i \leqslant p - K, \\ \theta_i & \text{if} \quad p - K + 1 \leqslant i \leqslant p. \end{cases} \tag{21}$$

We have to prove that

$$\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}\mathbf{z}_i = \frac{\mathbf{z}_i}{\gamma_i}$$

or equivalently

$$\gamma_i B^{1/2}\widehat{\Sigma}^{-1}B^{1/2}\mathbf{c}_i = \mathbf{c}_i.$$

This is the same as

$$B^{-1/2}\widehat{\Sigma}B^{-1/2}\mathbf{c}_i = \gamma_i\mathbf{c}_i$$

and by part (a) we only need to prove

$$B^{-1/2}(W + B^{1/2}C_{p-K}C'_{p-K}B^{1/2})B^{-1/2}\mathbf{c}_i = \gamma_i\mathbf{c}_i$$

which is the same as

$$(A + C_{p-K}C'_{p-K})\mathbf{c}_i = \gamma_i\mathbf{c}_i. \tag{22}$$

Since $A\mathbf{c}_i = \theta_i\mathbf{c}_i$ and $C_{p-K}C'_{p-K}\mathbf{c}_i = \mathbf{c}_i$ if $1 \leqslant i \leqslant p - K$ and $\mathbf{0}$ if $p - K + 1 \leqslant i \leqslant p$, (22) holds.

Since we have shown that $\mathbf{z}_i$ and $\widehat{\mathbf{t}}_{p-i+1}$ are both eigenvectors of the same matrix, associated with the same eigenvalue, to prove (c) it is enough to show that $\mathbf{v}_i = \widehat{\Sigma}^{1/2}\mathbf{z}_i$ is an eigenvector of BW^{-1} corresponding to the eigenvalue $1/\theta_i$. Then we have to prove that

$$BW^{-1}B^{1/2}\mathbf{c}_i = \frac{1}{\theta_i}B^{1/2}\mathbf{c}_i$$

and this is equivalent to

$$B^{1/2}W^{-1}B^{1/2}\mathbf{c}_i = \frac{1}{\theta_i}\mathbf{c}_i$$

which is true.

Proof of (d): Note that given $\widehat{\boldsymbol{\mu}}_h$, $1 \leqslant h \leqslant H$, the maximum likelihood estimate of Σ is the maximum likelihood estimate of Σ when we want to fit a normal $N_p(\mathbf{0}, \Sigma)$ distribution to the observations $\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h$, $1 \leqslant j \leqslant n_h$, $1 \leqslant h \leqslant H$.

Finally, to prove (6) note that

$$
\begin{aligned}
p &= \operatorname{trace}(I_p) = \operatorname{trace}(\widehat{\Sigma}^{-1}\widehat{\Sigma}) \\
&= \operatorname{trace}\left(\widehat{\Sigma}^{-1}\left(\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\right)\right) \\
&= \operatorname{trace}\left(\frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\widehat{\Sigma}^{-1}\right) \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\operatorname{trace}((\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\widehat{\Sigma}^{-1}) \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}\operatorname{trace}((\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\widehat{\Sigma}^{-1}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)) \\
&= \frac{1}{N}\sum_{h=1}^{H}\sum_{j=1}^{n_h}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h)'\widehat{\Sigma}^{-1}(\mathbf{x}_{hj} - \widehat{\boldsymbol{\mu}}_h). \quad \square
\end{aligned}
$$

Proof of Theorem 2. The vectors $\widehat{\boldsymbol{\eta}}_i$, $1 \leqslant i \leqslant K$, computed in step 5 of the SIR procedure are the eigenvectors corresponding to the K largest eigenvalues of $\widehat{\Psi}^{-1/2}B\widehat{\Psi}^{-1/2}$ and satisfy

$$
\widehat{\Psi}^{-1/2}B\widehat{\Psi}^{-1/2}\widehat{\boldsymbol{\eta}}_i = \omega_i\widehat{\boldsymbol{\eta}}_i, \tag{23}
$$

where $1 > \omega_1 \geqslant \cdots \geqslant \omega_p > 0$.

The estimate of the CMS obtained using the SIR algorithm is spanned by $\widehat{\boldsymbol{\beta}}_i = \widehat{\Psi}^{-1/2}\widehat{\boldsymbol{\eta}}_i$, $1 \leqslant i \leqslant K$. Eq. (23) is equivalent to

$$
B\widehat{\boldsymbol{\beta}}_i = \omega_i\widehat{\Psi}\widehat{\boldsymbol{\beta}}_i
$$

and using that $\widehat{\Psi} = B + W$ we get

$$
B\widehat{\boldsymbol{\beta}}_i(1 - \omega_i) = \omega_i W\widehat{\boldsymbol{\beta}}_i.
$$

Then, we have

$$
\widehat{\Psi}\widehat{\boldsymbol{\beta}}_i = \frac{1}{\omega_i}B\widehat{\boldsymbol{\beta}}_i = \frac{1}{1 - \omega_i}W\widehat{\boldsymbol{\beta}}_i, \tag{24}
$$

and then, by (24), we have

$$
BW^{-1}\widehat{\Psi}\widehat{\boldsymbol{\beta}}_i = BW^{-1}\frac{1}{1 - \omega_i}W\widehat{\boldsymbol{\beta}}_i = \frac{1}{1 - \omega_i}B\widehat{\boldsymbol{\beta}}_i = \frac{\omega_i}{1 - \omega_i}\widehat{\Psi}\widehat{\boldsymbol{\beta}}_i.
$$

Then $\widehat{\Psi}\widehat{\boldsymbol{\beta}}_i$ is an eigenvector of BW^{-1} corresponding to the eigenvalue $\omega_i/(1 - \omega_i)$, which is the i-th largest eigenvalue of BW^{-1}. By part (c) of Theorem 1, it turns out that $\widehat{\Sigma}^{1/2}\widehat{\mathbf{t}}_i$ is an eigenvector of BW^{-1} corresponding to its i-th largest eigenvalue. Then the subspace generated by $\widehat{\boldsymbol{\beta}}_i$, $1 \leqslant i \leqslant K$, is the same as the subspace generated by $\widehat{\Psi}^{-1}\widehat{\Sigma}^{1/2}\widehat{\mathbf{t}}_i$, $1 \leqslant i \leqslant K$, which is equal to the subspace $\widehat{\Psi}^{-1}\widehat{\mathscr{V}}^*$, where $\widehat{\mathscr{V}}^*$ is as in part (b) of Theorem 1. $\square$

References

Cook, R.D., 1994. On the interpretation of regression plots. Journal of the American Statistical Association 89, 177–189.

Cook, R.D., 1996. Graphics for regressions with a binary response. Journal of the American Statistical Association 91, 983–992.

Cook, R.D., 1998. Principal Hessian directions revisited. Journal of the American Statistical Association 93, 84–94.

Cook, R.D., 2007. Fisher lecture: dimension reduction in regression. Statistical Science 22, 1–26.

Cook, R.D., Li, B., 2002. Dimension reduction for conditional mean in regression. Annals of Statistics 30, 455–474.

Cook, R.D., Ni, L., 2005. Sufficient dimension reduction via inverse regression: a minimum discrepancy approach. Journal of the American Statistical Association 100, 410–428.

Cook, R.D., Weisberg, S., 1991. Discussion of sliced inverse regression for dimension reduction, by K.C. Li. Journal of the American Statistical Association 86, 328–332.

Cook, R.D., Wetzel, N., 1993. Exploring regression structure with graphics (with discussion). Test 2, 33–100.
Ferré, L., 1998. Determining the dimension in sliced inverse regression and related methods. Journal of the American Statistical Association 93, 132–140.
Li, K.C., 1991. Sliced inverse regression for dimension reduction. Journal of the American Statistical Association 86, 316–327.
Li, B., Zha, H., Chiaromonte, F., 2005. Contour regression: a general approach to dimension reduction. The Annals of Statistics 33, 1580–1616.
Li, K.-C., 1992. On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. Journal of the American Statistical Association 87, 1025–1039.
Schott, J.R., 1994. Determining the dimensionality in sliced inverse regression. Journal of the American Statistical Association 89, 141–148.
Seber, G.A.F., 1986. Multivariate Observations. Wiley, New York.
Zhu, Y., Zeng, P., 2006. Fourier methods for estimating the central subspace and the central mean subspace in regression. Journal of the American Statistical Association 101, 1638–1651.