A mixture model for dimension reduction

Jean-Luc Dortet-Bernadet, Laurent Gardes

▶ **To cite this version:**

Jean-Luc Dortet-Bernadet, Laurent Gardes. A mixture model for dimension reduction. Communications in Statistics - Theory and Methods, 2017, 46 (21), pp.10768-10787. 10.1080/03610926.2016.1248576 . hal-01077146v4

HAL Id: hal-01077146
https://hal.science/hal-01077146v4

Submitted on 27 Sep 2016

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 mixture model for dimension reduction

Jean-Luc Dortet-Bernadet and Laurent Gardes

Université de Strasbourg & CNRS, IRMA,

UMR 7501, 7 rue René Descartes,

67084 Strasbourg cedex, France.

Abstract

The existence of a Dimension Reduction (DR) subspace is a common assumption in regression analysis when dealing with high-dimensional predictors. The estimation of such a DR subspace has received considerable attention in the past few years, the most popular method being undoubtedly the Sliced Inverse Regression. We propose in this paper a new estimation procedure of the DR subspace by assuming that the joint distribution of the predictor and the response variables is a finite mixture of distributions. The new method is compared through a simulation study to some classical methods.

Keywords: Dimension reduction, Maximum likelihood estimates, Mixture of distributions, Sliced Inverse Regression.

1 Introduction

Regression analysis concerns inference on the conditional distribution of a response variable $Y \in \mathbb{R}^q$ given the value $X = x$ of a vector of predictors $X \in \mathbb{R}^p$. For instance, a classical problem is the nonparametric estimation of the conditional mean function $\mathbb{E}(Y|X)$ for which a popular estimator, when the dimension p is not too large, has been proposed by Nadaraya [22] and Watson [30]. When the dimension p becomes large, the so-called "curse of dimensionality" problem arises and

inference on the conditional distribution of Y given $X = x$ becomes difficult. A common procedure when dealing with a high-dimensional predictor X is to determine a subspace $\mathcal{S} \subset \mathbb{R}^p$, with $\dim(\mathcal{S}) = d \leq p$, that carries all the information that X has about Y. Such a subspace $\mathcal{S}$ is called a Dimension Reduction (DR) subspace. It is spanned by the columns of a full rank matrix $\Gamma \in \mathbb{R}^{p \times d}$ such that X and Y are conditionally independent given $\Gamma^t X$. A DR subspace always exists since the trivial choice $\Gamma = I_p$ is possible, but does not produce a reduction of dimension. Under minor conditions (see Cook [6]), the intersection of two DR subspaces is still a DR subspace and the intersection of all DR subspaces is called the central subspace. As seen in Li [19], a regression model admitting a central subspace is given by $Y = g(\Gamma^t X, \varepsilon)$, where ε is a random value independent of X and $g : \mathbb{R}^{d+1} \mapsto \mathbb{R}^q$ is an arbitrary function.

One of the earliest method to estimate the central subspace (*i.e.* a matrix Γ) is the Sliced Inverse Regression (SIR) procedure introduced by Li [19]. This method is based on the estimation of $\mathrm{Var}(\mathbb{E}(X|Y))$ using a set $\{S_h,\ h = 1, \ldots, H\}$ of non-overlapping slices that cover the range of Y. The asymptotic properties of SIR and related methods are derived for instance by Saracco [24, 25]. The SIR central subspace estimator is motivated in Li [19] by a geometric property of the covariance matrix $\mathrm{Var}(\mathbb{E}(X|Y))$. Another way to understand the SIR method is proposed in Cook [8] where Γ is interpreted as a parameter of an inverse regression model. This model is equivalent to assume that, for all $h = 1, \ldots, H$, the conditional distribution of X given $Y \in S_h$ is a multivariate Gaussian distribution. Considering n independent replications of the random vector (X, Y), Bernard-Michel *et al.* [3] and Szretter and Yohai [28] show that the maximum likelihood estimator of Γ corresponds to the SIR estimator of the central subspace. This inverse regression model is also used by Bernard-Michel *et al.* [3] to propose a Gaussian regularized version of SIR which is applied to a real data set in Bernard-Michel *et al.* [4].

The situation when the response variable Y is multivariate (*i.e.* when $dim(Y) = q > 1$) has received less attention in the literature. The main difficulty of applying SIR in this setting lies in

the construction of the non-overlapping slices. However, some adaptions of SIR to a multivariate framework have been proposed. For instance, a multivariate version of SIR where slices are replaced by k-means clusters is proposed in Setodji and Cook [27]. Hsing [15] describes a version of SIR for which the slices are built using a nearest neighbors approach. Yin and Bura [34] propose a moment based dimension reduction for multivariate data. More recently, Coudret *et al.* [11] present another extension of SIR that clusters components of a multivariate response variable Y that are related to the same DR subspace.

It is also well known that dimension reduction methods based on the first moment (as it is the case with the SIR method) fail to recover a symmetric dependency. This situation occurs for instance when the link function g in the Li's regression model is symmetric (see Cook and Weisberg [10]). A first tentative to overcome this limitation is proposed in Hsing and Carroll [16] who estimate the central subspace using an estimator of $\mathbb{E}(\mathrm{Var}(X|Y))$ instead of $\mathrm{Var}(\mathbb{E}(X|Y))$. One can also mention the following methods: Sliced Average Variance Estimation (SAVE) which is based on the second order moment of the conditional distribution of X given Y (see Cook [10]), Principal Hessian Directions (pHd) (Li [20]), Graphical Regression (Cook [7]), Minimum Average Variance Estimation (MAVE) (Xia *et al.* [32], Directional Regression (Li and Wang [17]), Sliced Regression (Wang and Xia [29]), Likelihood Acquired Directions (LAD) (Cook and Forzani [9]) and many others. Convex combinations of some of the previous methods are investigated in Gannoun and Saracco [13] and Ye and Weiss [33]. Note that most of these methods have been introduced for the case of a univariate response variable Y. A few extensions to the multivariate case can be found in Aragon [2] and Li *et al.* [21].

The goal of this paper is to propose a new dimension reduction approach. In few words, we assume that the whole joint distribution of (X, Y) is a finite mixture of distributions, parametrized in such a way that it allows inference on the central subspace. The proposed method avoids the choice of non-overlapping slices and is thus well adapted to the presence of a multivariate response

variable Y. Moreover the proposed method is able to recover a symmetric dependency. Notice that the use of models based on mixtures of distributions has been already proposed in the context of dimension reduction, only for a univariate response Y. For example, Scrucca [26] assumes that the conditional distribution of X given $Y \in S_h$ is a finite mixture of Gaussian distributions. Reich *et al.* [23], in a Bayesian framework, propose a mixture model for the conditional distribution of a real-valued response Y given X with a probit model on the weights.

The rest of the paper is organized as follows. In Section 2, the new dimension reduction model is introduced and an estimation of its parameters is provided. A comparison with existing methods is given in Section 3. A simulation study is proposed in Section 4 where our new estimation procedure is compared to previous approaches. A real dataset is treated in Section 5. All the proofs are postponed to the appendix.

2 The Dimension reduction estimation procedure

2.1 The proposed model

We assume in what follows that the random vector $(X, Y) \in \mathbb{R}^p \times \mathbb{R}^q$ admits a probability density function (pdf) $f_{X,Y}(x, y)$ with respect to the Lebesgue measure. Our aim is to estimate a full rank matrix $\Gamma \in \mathbb{R}^{p\times d}$ such that the columns of Γ form a basis of a DR subspace of dimension $d \leq p$. For that purpose, for an integer $M \geq d + 1$, we suppose that the joint distribution of (X, Y) is a mixture of M distributions involving Γ. More specifically, for some unknown positive component weights $\pi_1, \ldots, \pi_M$ summing to 1 we state that

$$f_{X,Y}(x, y) = \sum_{m=1}^{M} \pi_m f_m(x, y|\Gamma), \tag{1}$$

where for each $m \in \{1, \ldots, M\}$, $f_m(\cdot, \cdot|\Gamma)$ is a pdf. To ensure that (1) is a dimension reduction model (*i.e.* that X and Y are conditionally independent given $\Gamma^t X$) we assume in addition that there exist functions (not necessarily pdf) $g(\cdot)$ and $h_m(\cdot, \cdot)$, $m = 1, \ldots, M$ such that

$$f_m(x, y|\Gamma) = g(x)h_m(\Gamma^t x, y). \tag{2}$$

Indeed, it is easy to check that under (1) with $f_m(\cdot,\cdot|\Gamma)$ as in (2), the pdf of the conditional distribution of Y given $\{X = x\}$ is

$$\begin{aligned} f(y|X=x) &= \sum_{m=1}^{M} \pi_m g(x) h_m(\Gamma^t x, y) \Bigg/ \left(g(x) \sum_{m=1}^{M} \pi_m \int h_m(\Gamma^t x, z) dz \right) \\ &= \sum_{m=1}^{M} \pi_m h_m(\Gamma^t x, y) \Bigg/ \sum_{m=1}^{M} \pi_m \int h_m(\Gamma^t x, z) dz \ , \end{aligned}$$

which depends on x only through $\Gamma^t x$. Notice that, as it is the case for all the dimension reduction models proposed in the literature, the matrix Γ is identifiable only up to a right product by any $d \times d$ regular matrix whereas the corresponding spanned subspace, the true goal of inference, is identifiable.

The advantage of assuming that the joint distribution of (X, Y) is a mixture distribution will clearly appear in Section 3 where a comparison with other dimension reduction methods is done.

Of course, without assuming a parametric form for the functions $\{f_m(\cdot,\cdot|\Gamma),\ m = 1, \ldots, M\}$, the estimation of Γ is impossible. We introduce now a natural example of parametric mixture distribution that will be used in the rest of the paper. We assume that, for each $m = 1, ..., M$, the pdf $f_m(\cdot,\cdot|\Gamma)$ is the product of a p-dimensional Gaussian pdf with mean $\xi + V\Gamma\beta_m \in \mathbb{R}^p$ (for $\xi \in \mathbb{R}^p$ and $\beta_m \in \mathbb{R}^d$) and covariance matrix $V \in \mathbb{R}^{p\times p}$ and a q-dimensional Gaussian pdf with mean $\alpha_m \in \mathbb{R}^q$ and covariance matrix $W_m \in \mathbb{R}^{q\times q}$. The distribution of (X, Y) is the Gaussian mixture

$$f_{X,Y}(x, y) = \sum_{m=1}^{M} \pi_m \varphi_p(x|\xi + V\Gamma\beta_m; V)\varphi_q(y|\alpha_m; W_m), \tag{3}$$

where $\varphi_k(\cdot|\mu; \Phi)$ denotes the pdf of a multivariate Gaussian distribution with mean $\mu \in \mathbb{R}^k$ and covariance matrix $\Phi \in \mathbb{R}^{k\times k}$. Note that for each $m \in \{1, \ldots, M\}$ the pdf of the m-th component $f_m(x, y|\Gamma) = \varphi_p(x|\xi + V\Gamma\beta_m; V)\varphi_q(y|\alpha_m; W_m)$ satisfies (2) with

$$g(x) = \frac{1}{(2\pi)^{p/2}|V|^{1/2}} \exp\left[-\frac{1}{2}(x-\xi)^t V^{-1}(x-\xi)\right] \tag{4}$$

for $x \in \mathbb{R}^p$ and, for $(x, y) \in \mathbb{R}^{p\times q}$,

$$h_m(\Gamma^t x, y) = \exp\left[\beta_m^t(\Gamma^t x - \Gamma^t\xi) - \frac{1}{2}\beta_m^t\Gamma^t V\Gamma\beta_m\right] \varphi_q(y|\alpha_m; W_m). \tag{5}$$

The expression for the conditional mean of X given that it comes from the m-th component distribution, $\xi + V\Gamma\beta_m$, is present in many works dedicated to dimension reduction. It is used *e.g.* in Cook [8] (Sec. 3.1) in a regression setting, for the case where V is the identity matrix. The general expression was introduced in Bernard-Michel *et al.* [3] to define their model of Gaussian sliced inverse regression.

The parameters of the Gaussian mixture model (3) are Γ (which is the parameter of interest), $\Theta = (\xi, V)$ and $\Theta_m = (\beta_m, \alpha_m, W_m)$ for $m = 1, \ldots, M$. It is of course possible to consider more parsimonious models. For instance, one can assume that the covariance matrix in $\varphi_q(y|\alpha_m; W_m)$ is given by $W_m = v^2 I_q$ for some real parameter $v > 0$. This parsimonious model will be used in the simulation study and for the real data example. Finally note that the extension of the model to other types of response variables, such as binary variables, is possible by replacing in (3) the pdf's $\varphi_q(y|\alpha_m; W_m)$, $m = 1, ..., M$ by appropriate distributions.

The next section is devoted to the estimation of the parameters involved in (3).

2.2 Maximum likelihood estimation

Let (X, Y) be a random vector with pdf given by (3). Let $(x, y) := ((x_1, y_1), \ldots, (x_n, y_n))$ be the observations of n independent copies of the random vector (X, Y). We propose to estimate the full rank matrix Γ spanning the DR subspace by its maximum likelihood estimator. Our goal is thus to maximize the likelihood function

$$\mathcal{L}\left((x, y)\,|\Gamma, \Theta, (\pi_m, \Theta_m)_{m=1,\ldots,M}\right) = \prod_{i=1}^{n} \sum_{m=1}^{M} \pi_m g(x_i) h_m(\Gamma^t x_i, y_i)$$

with respect to Γ, Θ, Θ_m and π_m, $m = 1, \ldots, M$ where the parametric functions $g(\cdot)$ and $\{h_m(\cdot, \cdot),\ m = 1, \ldots, M\}$ are defined in (4) and (5). To achieve this maximization we use the Expectation-Maximization (EM) algorithm (see Dempster *et al.* [12]). The idea behind this algorithm is the following. We introduce a latent variable Z taking values in $\{1, \ldots, M\}$ with $\mathbb{P}(Z = m) = \pi_m$ and such that the conditional pdf of (X, Y) given $Z = m$ is $f_m(x, y|\Gamma) = g(x)h_m(\Gamma^t x, y)$. The EM algorithm is an iterative procedure maximizing the expectation of the complete log-likelihood, *i.e* the log-likelihood of the random vector (X, Y, Z).

Note that for the Gaussian mixture model (3), a problem related to the identifiability of the parameters arises: for any $\gamma \in \mathbb{R}^d$, the distribution is unchanged by the reparameterization $\tilde{\xi} = \xi + V\Gamma\gamma$ and $\tilde{\beta}_m = \beta_m - \gamma$ since $\xi + V\Gamma\beta_m = \tilde{\xi} + V\Gamma\tilde{\beta}_m$. To overcome this problem, it is assumed in what follows that $\beta_M = 0$.

To describe the estimator of the DR subspace provided by the EM algorithm we first introduce the following notations: let

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{and} \quad \hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^t$$

be the empirical mean and variance matrix of X. For $i = 1, \ldots, n$ and $m = 1, \ldots, M$, let $z_{i,m}$ be the estimator of $\mathbb{P}(Z = m|(X, Y) = (x_i, y_i))$ provided by the EM algorithm and let $\hat{C}_n$ be the $p \times p$ matrix defined by

$$\hat{C}_n = \sum_{m=1}^{M} \hat{\pi}_m(\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^t,$$

where

$$\hat{\pi}_m = \frac{1}{n}\sum_{i=1}^{n} z_{i,m} \quad \text{and} \quad \bar{x}_m = \frac{1}{n\hat{\pi}_m}\sum_{i=1}^{n} z_{i,m}x_i.$$

The expression for $z_{i,m}$ is given in Lemma 1 of the Appendix, where it is also shown that $\hat{\pi}_m$ is the maximum likelihood estimator of π_m. Furthermore, $\bar{x}_m$ can be interpreted as an estimator of $\mathbb{E}(X|Z = m)$ and $\hat{C}_n$ as an estimator of the variance matrix $C := \mathrm{Var}(\mathbb{E}(X|Z))$.

Theorem 1. *Assume that model (3) holds. The maximum likelihood estimator of the DR subspace $\mathcal{S}_\Gamma$ is the space spanned by the d eigenvectors corresponding to the d largest eigenvalues of the matrix $\hat{\Sigma}_n^{-1}\hat{C}_n$.*

Roughly speaking, Theorem 1 entails that the maximum likelihood estimator of Γ maximizes the between group variance of X where the groups are the M latent classes.

The EM algorithm is detailed in the Appendix. In particular, Proposition 1 in Appendix gives all the maximum likelihood estimates of the parameters of model (3) required by the M-step of the algorithm.

3 Comparison with other models

According to Theorem 1, the proposed estimator of the DR subspace is based on a spectral decomposition of an estimator of the $p \times p$ matrix $\Sigma^{-1}C$ where $\Sigma = \mathrm{Var}(X)$ and $C = \mathrm{Var}(\mathbb{E}(X|Z))$. In fact many reduction methods are based on a spectral decomposition of a matrix. This is the case for example of the classical SIR method of Li [19] and of the more recent MSIR method of Scrucca [26] that uses mixtures of distributions. We give here some details on the similarities and the differences between these two methods and the proposed method.

SIR approach As shown for instance in Bernard-Michel *et al.* [3], for a univariate response Y, the estimate of Γ obtained by the SIR procedure of Li [19] maximizes the likelihood function of the model given by

$$f_{X,Y}(x,y) = \sum_{h=1}^{H} \mathbb{I}_{\{y \in S_h\}} \varphi_p(x|\xi + V\Gamma\beta_h; V) f_Y(y), \tag{6}$$

where $f_Y(\cdot)$ is an arbitrary pdf function and where $\{S_h,\ h = 1, \ldots, H\}$ are non-overlapping slices covering the range of Y. These slices have to be chosen by the user on the only basis of the observed distribution of Y. The SIR method estimates Γ by a spectral decomposition of an estimator of the $p \times p$ matrix $\Sigma^{-1}C^{(SIR)}$ where the matrix $C^{(SIR)} = \mathrm{Var}(\mathbb{E}(X|Y))$ is estimated by

$$\hat{C}_n^{(SIR)} = \sum_{h=1}^{H} \frac{n_h}{n} \left(\left(\frac{1}{n_h} \sum_{i:Y_i \in S_h} x_i \right) - \bar{x} \right) \left(\left(\frac{1}{n_h} \sum_{i:Y_i \in S_h} x_i \right) - \bar{x} \right)^t,$$

with n_h the number of observed $Y_i's$ in slice S_h, $h = 1, \ldots, H$. Hence, the SIR estimator of Γ is obtained by maximizing the between group variance of X where the groups are the H non-overlapping slices $\{S_h,\ h = 1, \ldots, H\}$.

MSIR approach A recent contribution to dimension reduction can be found in Scrucca [26] that describes a model-based SIR (MSIR) procedure. The idea here is to replace each Gaussian component in model (6) by a mixture of Gaussian distributions in order to deal with more complex situations. For a univariate response Y and for non-overlapping slices $\{S_h,\ h = 1, \ldots, H\}$, the model that is considered is

$$f_{X,Y}(x,y) = \sum_{h=1}^{H} \mathbb{I}_{\{y \in S_h\}} \sum_{j=1}^{J_h} q_{h,j} \varphi_p(x|\mu_{h,j}; \Sigma_{h,j}) f_Y(y) \tag{7}$$

where, for each $h = 1, \ldots, H$, the reals $\{q_{h,j},\ j = 1, \ldots, J_h\}$ are summing to 1, the vectors $\mu_{h,j} \in \mathbb{R}^p$ and the matrices $\Sigma_{h,j} \in \mathbb{R}^{p \times p}$, $j = 1, \ldots, J_h$, are unknown parameters. The DR subspace is defined as the space spanned by the d eigenvectors associated to the largest eigenvalues of the matrix $\Sigma^{-1} C^{(MSIR)}$ where $C^{(MSIR)} = \mathrm{Var}(\mathbb{E}(X|Y, Z^*))$, Z^* being the latent variable giving the mixture components in the slices. This matrix $C^{(MSIR)}$ is estimated by

$$\hat{C}_n^{(MSIR)} = \sum_{h=1}^{H} \sum_{j=1}^{J_h} \frac{n_h}{n} \hat{q}_{h,j} \left(\hat{\mu}_{h,j} - \hat{\mu}\right) \left(\hat{\mu}_{h,j} - \hat{\mu}\right)^t$$

with

$$\hat{\mu} = \sum_{h=1}^{H} \sum_{j=1}^{J_h} \frac{n_h}{n} \hat{q}_{h,j} \hat{\mu}_{h,j}$$

and where, for $h = 1, \ldots, H$ and $j = 1, \ldots, J_h$, $\hat{q}_{h,j}$ and $\hat{\mu}_{h,j}$ are obtained by fitting the mixture model (7). As the SIR estimator, the MSIR estimator of Γ is thus obtained by maximizing a between group variance of X but when the groups are the $J_1 + \ldots + J_H$ classes $\{S_h \times T_j^{(h)},\ j = 1, \ldots, J_h,\ h = 1, \ldots, H\}$ where $\{T_j^{(h)},\ j = 1, \ldots, J_h\}$ are the latent classes in the slice S_h.

Figure 1 illustrates the behavior of the SIR, MSIR and proposed methods on a simple example. Here the DR subspace has dimension $d = 1$ and $n = 200$ observations of a real response variable Y are simulated from the model $Y = (\Gamma^t X)^2 + \varepsilon$ for $p = 2$, $\Gamma^t = (1, 1)$, for values of X sampled from a standard Gaussian distribution and for $\varepsilon \sim \mathcal{N}(0, 1.5^2)$. The SIR method considers $\mathbb{E}(X|Y)$ and approximates this conditional expectation by fitting within each slice S_h, $h = 1, ..., H$, a single Gaussian distribution on the observations of X. Thus a well known drawback of this method is that it is inefficient when Y is a symmetric function of $\Gamma^t X$, as here, since in this case $\mathbb{E}(X|Y)$ is constant. The MSIR approach uses a mixture of distributions within each slice. The corresponding model is thus clearly more flexible than the SIR model (6), in particular can handle the case of a symmetric dependency. But it still depends on a choice of non-overlapping slices and thus it is difficult to adapt this method to multivariate response variables. Notice also that the MSIR estimator for Γ proposed by Scrucca [26] cannot be interpreted as a maximum likelihood estimator. By contrast with these methods, the proposed procedure does not need pre-defined slices and is

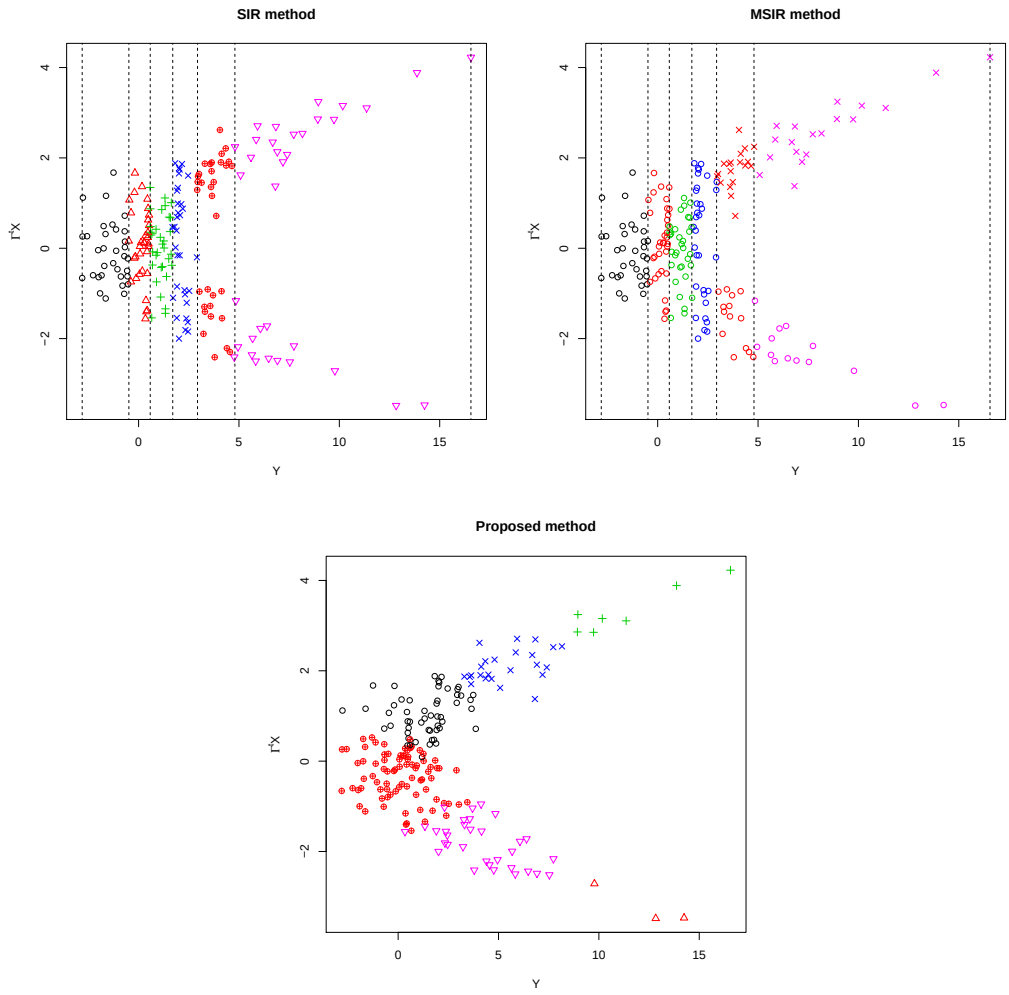

Figure 1: Plot of observed values of $(Y, \Gamma^t X)$ for the model $Y = (\Gamma^t X)^2 + \varepsilon$. The SIR method uses slices on the range of Y (delimited by dashed lines), here constructed from empirical quantiles of the observations. The MSIR method allows to use a mixture of distributions within each slice (the different symbols correspond to the classification given by the use of the mixture of distributions). The proposed method does not need slices and uses a mixture of distributions for (X, Y).

fully data-driven. The slices are replaced by the M latent classes of the mixture model (1) which are estimated by the EM algorithm. As mentioned in the introduction, this is the main motivation of using a mixture model for the joint distribution of (X, Y). As a consequence, the proposed method is well adapted to the case of multivariate response variables and can tackle complex situations like for instance a symmetric relationship between Y and $\Gamma^t X$.

4 Simulation study

In this section we examine the performance of the proposed method *via* a simulation study. The algorithm used here corresponds to model (3) where the conditional distribution of Y given that it comes from component m of the mixture is Gaussian with mean α_m and common covariance matrix $W_m = v^2 I_q$. The unknown parameters of this parsimonious model are the matrices $\Gamma \in \mathbb{R}^{p\times d}$ and $V \in \mathbb{R}^{p\times p}$, the vector $\xi \in \mathbb{R}^p$, the scalar $v^2 > 0$, the $M-1$ vectors β_m in $\mathbb{R}^d$ and the M vectors α_m in $\mathbb{R}^q$.

In practice, to run the EM algorithm, a starting value for each quantity $z_{i,m}$ is needed. To avoid local maxima and to get a more precise result, several starting values are used, retaining the estimation returned by the algorithm with the highest likelihood. Several ways are possible to define these different starting values. We have considered hierarchical clustering with different agglomeration methods and projection of the x_i's on the DR subspace provided by the SIR or SAVE methods.

We list below the simulation designs that are used along this simulation study. Notice that they are classical for the study of a dimension reduction method, the datasets are not designed to fit the proposed model.

• **Univariate case** ($q = 1$): let X be a standard Gaussian random vector of dimension p and let ε be a random value independent of X and following a normal distribution with mean 0 and standard deviation 0.2. For a given full rank matrix $\Gamma \in \mathbb{R}^{p\times d}$ and for a function $\mathcal{G} : \mathbb{R}^{d+1} \mapsto \mathbb{R}$, the response variable Y is given by the model $Y = \mathcal{G}(\Gamma^t X, \varepsilon)$. In the following designs, the matrix

Γ and the functions $\mathcal{G}$ are taken as in Li and Wang [17]. More precisely, we take $p = 6$, $d = 2$,

$$\Gamma^t = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 3 \end{pmatrix},$$

with $\mathcal{G}(\Gamma^t x, \varepsilon)$ given by, if $(\gamma_1, \gamma_2) = \Gamma^t x$,

$$\text{Design 1: } 0.4\gamma_1^2 + 3\sin(\gamma_2/4) + \varepsilon \quad \text{Design 2: } 3[\sin(\gamma_1/4) + \sin(\gamma_2/4)] + \varepsilon$$

$$\text{Design 3: } 0.4\gamma_1^2 + |\gamma_2|^{1/2} + \varepsilon \qquad \text{Design 4: } 3\sin(\gamma_2/4) + (1+\gamma_1)^2\varepsilon.$$

The next design is considered for instance in Li [19]. We take $p = 10$, $d = 2$,

$$\Gamma^t = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \end{pmatrix}, \tag{8}$$

with $\mathcal{G}(\Gamma^t x, \varepsilon)$ given by, if $(\gamma_1, \gamma_2) = \Gamma^t x$,

$$\text{Design 5: } 3\gamma_1 \left[0.5 + (\gamma_2 + 1.5)^2\right]^{-1} + \varepsilon.$$

• **Multivariate case** ($q > 1$): To study the performance of the proposed method when the response Y is multivariate we consider four simulation designs used in Setodji and Cook [27]. Here again, X is a standard Gaussian random vector of dimension p and ε is a standard Gaussian vector of dimension q independent of X. The response variable Y is given by $\mathcal{G}(\Gamma^t X, \epsilon)$ where $\mathcal{G} : \mathbb{R}^{d+q} \mapsto \mathbb{R}^q$. In the two following designs, we take $p = 4$, $q = 4$, $d = 1$, $\Gamma^t = (1, 1, 1, 1)$ with $\mathcal{G}(\Gamma^t x, \varepsilon)$ given by, if $\gamma = \Gamma^t x$ and $\varepsilon = (\varepsilon_1, \varepsilon_2, \varepsilon_3, \varepsilon_4)$,

$$\text{Design 6: } \frac{\gamma}{10} + \left(\varepsilon_1 \exp\left(\frac{\gamma}{10}\right),\ \varepsilon_2 \exp\left(\frac{2-3\gamma}{10}\right),\ \varepsilon_3 \exp\left(\frac{\gamma}{5}\right),\ \varepsilon_4 \exp\left(\frac{1-\gamma}{10}\right)\right).$$

$$\text{Design 7: } \left(\varepsilon_1 \exp\left(\frac{\gamma}{10}\right),\ \varepsilon_2 \exp\left(\frac{2-3\gamma}{10}\right),\ \varepsilon_3 \exp\left(\frac{\gamma}{5}\right),\ \varepsilon_4 \exp\left(\frac{1-\gamma}{10}\right)\right).$$

For the last simulation designs, we take $p = 10$, $q = 2$, $d = 2$, the matrix Γ as in (8) and with $\mathcal{G}(\Gamma^t x, \varepsilon)$ given by, if $(\gamma_1, \gamma_2) = \Gamma^t x$ and $\varepsilon = (\varepsilon_1, \varepsilon_2)$,

$$\text{Designs 8 and 9: } \left(\gamma_1(\gamma_1 + \gamma_2 + 1) + \sigma\varepsilon_1,\ \gamma_1[0.5 + (\gamma_2 + 1.5)^2]^{-1} + \sigma\varepsilon_2\right),$$

with $\sigma = 1/2$ for Design 8 and $\sigma = 1$ for Design 9.

Let P be the matrix of projection on the true DR subspace and $\widehat{P}$ the matrix of projection on the estimated DR subspace. Following Li *et al.* [18] and Scrucca [26], to measure the accuracy of the proposed method, we calculate the Euclidean norm of $P - \widehat{P}$ which is defined as the maximum singular value of $(P-\widehat{P})(P-\widehat{P})$. This norm have values in the interval $(0, 1)$ and can be interpreted as a sine of the maximal angle between the true and the estimated DR subspaces. The results presented hereafter are calculated over 100 data replications.

4.1 Choice of M

We first examine how the proposed method performs depending on the choice for the number M of components in the mixture. For this, we run the algorithm for different data size values and for values of M going from 3 to 40 components.

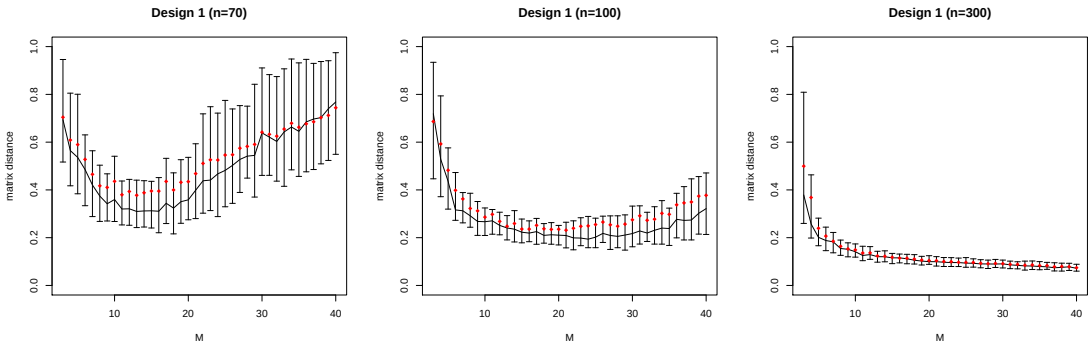

Figure 2: Choice of M. Design 1: Quantiles of order 0.25, 0.5 and 0.75 (solid lines) and means (diamonds) of the matrix distances as a function of M for data sizes $n = 70, 100$ and 300.

In Figure 2 we report for the design 1 the quartiles and the means of the matrix distances for three data sizes $n = 70, 100$ and 300. The errors decrease with M until they reach a plateau. Then, as seen in the case $n = 70$ and $n = 100$, the results deteriorate when the number of components is too large for the data size. The results look insensitive to the choice of a "reasonable" value for M, reasonable value that is sufficiently large to apprehend the link function and not too big with respect to n and the values for p, q and d.

In practice, since the true DR subspace is not known, one can look for a reasonable M by checking

the stability of the estimates. Another possibility is to do a cross-validation to verify if the model recovers well the link function. Nevertheless such practices are computationally expensive. Hereafter we will use in this section the choice by default $M = \lfloor 2n^{0.5} \rfloor$ that appears to work well for the simulations.

4.2 Comparison with other methods

Next we compare the performance of the proposed method with other methods for dimension reduction on the different simulation designs. For the case of a univariate response variable we consider the classical methods SIR and SAVE that are implemented in the R package dr (see Weisberg [31]), the method LAD of Cook and Forzani [9] that is implemented in the R package ldr (see Adragni and Raim [1]) and the method MSIR implemented in the R package msir (see Scrucca [26]).

		Our method	SIR	SAVE	LAD	MSIR
Design 1	$n = 100$	0.203 (0.070)	0.765 (0.212)	0.425 (0.164)	0.340 (0.169)	0.533 (0.300)
	$n = 300$	0.085 (0.027)	0.675 (0.267)	0.181 (0.064)	0.149 (0.049)	0.171 (0.081)
Design 2	$n = 100$	0.745 (0.230)	0.814 (0.196)	0.819 (0.173)	0.727 (0.227)	0.853 (0.153)
	$n = 300$	0.364 (0.194)	0.663 (0.213)	0.710 (0.220)	0.335 (0.173)	0.721 (0.210)
Design 3	$n = 100$	0.440 (0.257)	0.926 (0.101)	0.382 (0.173)	0.272 (0.155)	0.844 (0.206)
	$n = 300$	0.108 (0.045)	0.915 (0.106)	0.158 (0.049)	0.107 (0.032)	0.783 (0.252)
Design 4	$n = 100$	0.743 (0.205)	0.839 (0.202)	0.787 (0.197)	0.755 (0.221)	0.902 (0.128)
	$n = 300$	0.300 (0.176)	0.795 (0.195)	0.526 (0.238)	0.420 (0.224)	0.900 (0.139)
Design 5	$n = 100$	0.779 (0.189)	0.686 (0.165)	0.952 (0.064)	0.828 (0.143)	0.782 (0.140)
	$n = 300$	0.376 (0.117)	0.391 (0.107)	0.874 (0.127)	0.459 (0.158)	0.454 (0.117)

Table 1: Mean and standard error (in parenthesis) of the matrix distances.

In Table 1 we report the means and standard errors of the matrix distances calculated over 100 data replications. The values of M for the proposed method are fixed by default ($M = 20$ for the

data size $n = 100$ and $M = 34$ for the data size $n = 300$). We run the SIR and SAVE methods, for each data replication, for a number of slices H between 3 and 40. For each of these methods we report the results corresponding to the number H that gives the smaller mean of the data distances. We proceed similarly for LAD, for a number of slices H between 3 and 12 when $n = 100$ and between 3 and 25 for $n = 300$ (fixing d to its true value). Finally the reported results for the MSIR method are the one obtained by the default value for H in the R function. Clearly, the proposed method performs very well compared to the other methods since it gives generally the best results, or a result close the the best one, except for design 3 when $n = 100$ and design 5 when $n = 100$ (even if the best result for SIR, SAVE and LAD over several number of slices are reported).

		Our method	SIR	SAVE
Design 6	$n = 100$	0.366 (0.176)	0.667 (0.242)	0.863 (0.171)
	$n = 300$	0.139 (0.067)	0.362 (0.205)	0.803 (0.199)
Design 7	$n = 100$	0.369 (0.139)	0.841 (0.176)	0.861 (0.165)
	$n = 300$	0.185 (0.082)	0.822 (0.186)	0.821 (0.191)
Design 8	$n = 200$	0.342 (0.112)	0.509 (0.142)	0.770 (0.188)
	$n = 400$	0.194 (0.056)	0.339 (0.087)	0.562 (0.189)
Design 9	$n = 200$	0.661 (0.197)	0.783 (0.174)	0.867(0.141)
	$n = 400$	0.401 (0.158)	0.596 (0.179)	0.697 (0.183)

Table 2: Mean and standard error (in parenthesis) of the matrix distances.

For the case $q > 1$ we compare the proposed method only with SIR and SAVE as LAD and MSIR consider only univariate response variables. The results are given in Table 2. The values of M are fixed by default ($M = 28$ for the data size $n = 200$ and $M = 40$ for the data size $n = 400$). Again, the reported values for SIR and SAVE correspond to the best among the results of these methods for a number of slices H between 3 and 15. Here again the proposed method compares very favorably to the others. This is particularly true for design 6 and design 7 where the high

dimension of Y, $q = 4$, makes difficult the construction of slices.

4.3 Choice of the dimension of the DR subspace

The estimation of the dimension d of a DR subspace is an important issue for a dimension-reduction method. To answer this question, many methods related to SIR use the sequential chi-squared test procedure introduced by Li [19] based on the test statistic

$$\Lambda_d = n \sum_{j=d+1}^{p} \hat{\lambda}_j$$

where the values $\hat{\lambda}_j$, $j = 1, ..., p$, denote the observed eigenvalues of the eigen decomposition of the estimator of $\Sigma^{-1}C^{(SIR)}$. Under some conditions on the distribution of X (see Bura and Cook [5]), concerning the SIR method, this statistic is known to have an asymptotic chi-squared distribution if d is the true dimension. For other methods than SIR, or when the conditions are not satisfied, the null distribution of this statistics can be explored *via* a Monte-Carlo-type procedure as in the general permutation test of Cook and Weisberg [10].

Such a method could be adapted to our case, but the Monte-Carlo study that needs many starts of the EM algorithm is computationally expensive in practice. We consider here the use of a simple sequential procedure based on the study of the decay of the eigen values. Let $\hat{\lambda}_j^{d_0}$, $j = 1, ..., p$, denote the observed eigen values of $\hat{\Sigma}_n^{-1}\hat{C}_n$ when d is fixed to d_0. As described in the proof of Proposition 1 in the Appendix section, these eigen values are such that $1 \geq \hat{\lambda}_1^{d_0} \geq \hat{\lambda}_2^{d_0} \geq ... \geq \hat{\lambda}_p^{d_0} \geq 0$. If the dimension d of the true DR subspace is greater than d_0 then $\hat{\lambda}_{d_0}^{d_0}$ is expected to be significantly much larger than $\hat{\lambda}_{d_0+1}^{d_0}, ..., \hat{\lambda}_p^{d_0}$. Comparing the observed eigen values with values equally spaced between 1 and 0, we can estimate the dimension of the DR subspace as

$$\hat{d} = \max \left\{ d : \hat{\lambda}_d^d \geq 1 - \frac{d}{p+1} \right\}. \tag{9}$$

This procedure is called EIV in the rest of the paper.

Another approach for dimension selection is to use an information criterion such as BIC or AIC. This has been considered in various works dedicated to dimension reduction that use a likelihood,

see *e.g.* Cook and Forzani [9]. For each d we calculate a penalized likelihood

$$\begin{aligned} BIC(d) &= \hat{L} - 0.5k(d)\log(n) \\ AIC(d) &= \hat{L} - k(d) \end{aligned}$$

where $\hat{L}$ denotes the maximum value of the likelihood and where $k(d)$ denotes the number of free parameters changing with d. Since in our model the parameters concerned by d are the matrix $\Gamma \in \mathbb{R}^{p\times d}$, identifiable only up to a right product by any regular matrix $D \in \mathbb{R}^{d\times d}$, and the $M-1$ vectors β_m in $\mathbb{R}^d$, we set $k(d) = d(p-d+M-1)$. For each criterion, the dimension selected is the d that returns the maximum value.

We evaluate the performances of these procedures on some simulations. We consider the designs 1, 3, 6 and 8 that give, according to Tables 1 and 2, a relatively good estimation of the DR subspace when d is known and that cover different situations. For each of these designs and for 5 data sizes, $n = 100, 200, 300, 400$ and $n = 500$, we calculate the number of times that each procedure estimates the correct dimension d over 100 data replications. We report the ratios of good answers in Figure 3. Shortly, the results are globally satisfactory since these ratios tend to grow with n. From these simulations there is not a clear ranking of these procedures of the estimation of d. The results of AIC and EIV look overall similar. If the procedure based on BIC can work very well in some situations, as for the design 6, it can be outclassed by the procedure based on AIC for relatively small data sizes, as seen with designs 3 and 8.

5 Real data

To illustrate the use of the proposed method in a multivariate context we consider the Minneapolis schools dataset. This dataset is described in Cook [7] and concerns the performance of students in $n = 63$ Minneapolis schools along with some various social and economic variables. It is studied, among others, in Yin and Bura [34] or more recently in Coudret *et al.* [11]. We follow these authors and consider a $q = 4$ dimensional response variable Y that consists of the percentages of students in a school scoring above and below average on standardized fourth and sixth grade reading com-

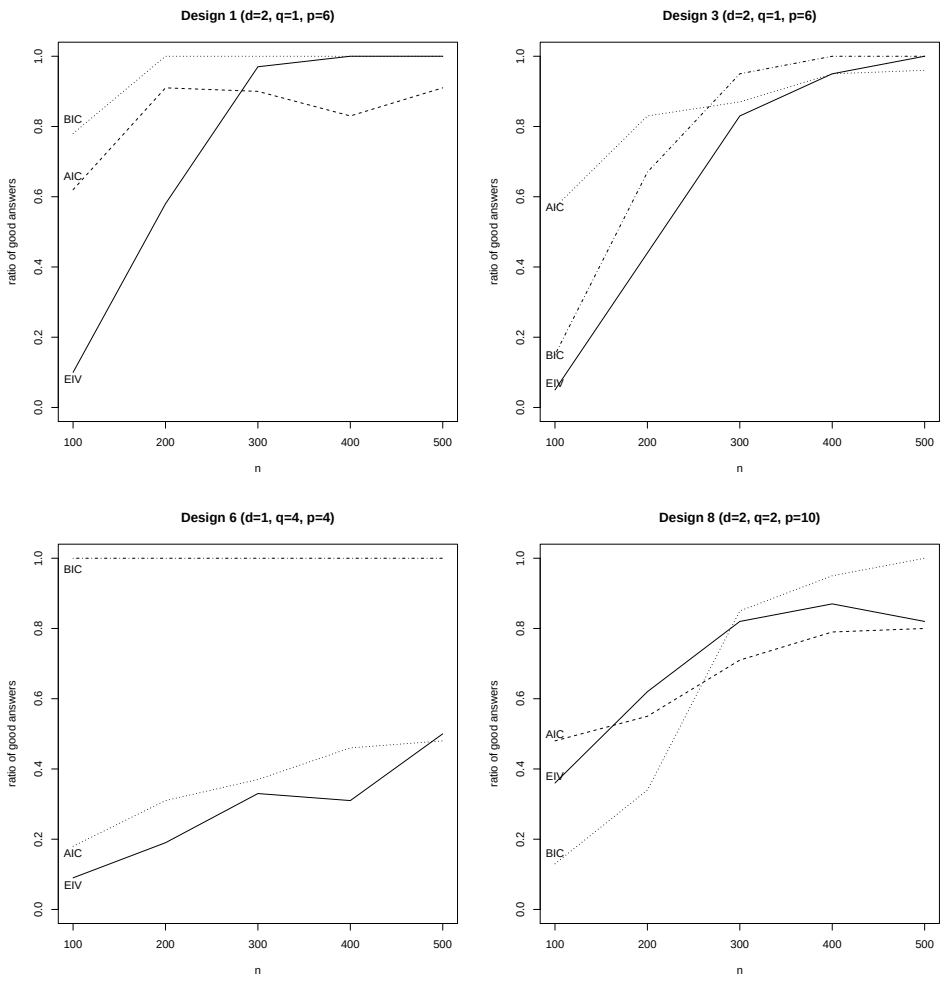

Figure 3: Choice of d. Ratio of good answers for different values of n for the procedure EIV, and the procedures based on AIC and BIC.

prehension tests, denoted Y4Below, Y4Above, Y6Below and Y6Above. As in Yin and Bura [34] the percentage of students scoring about average is not used since the sum of this with the percentages above and below average is 100%. There are $p = 8$ potential predictors. The first seven are the squared root of the percentages of: children receiving an aid called AFDC, children who do not live with both parents, people in the area of a school who completed high school, people who suffer for poverty, minority, mobility and pupils who attend school regularly. The eighth predictor is the mean number of pupils for each teacher. Notice that the small data size and the dimension of Y makes the construction of slices tricky: the SIR and SAVE methods implemented in the package

dr are ineffective in this situation.

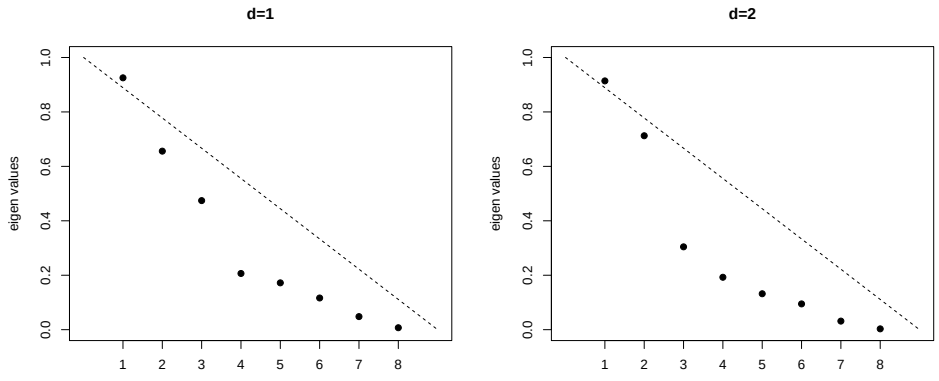

Figure 4: Minneapolis schools dataset. Eigen values obtained for $d = 1$ and $d = 2$.

The choice by default $M = \lfloor 2n^{0.5} \rfloor$ used in the simulations of the previous section gives here $M = 15$. This number being maybe too large for a so small data size, $n = 63$, we report hereafter the results obtained for the choice $M = 10$ (the results obtained for values of M around 10 are very similar).

We first look at the dimension of the DR subspace. We report in Figure 4 the eigen values of $\hat{\Sigma}_n^{-1}\hat{C}_n$ for $d = 1$ and $d = 2$. Clearly, going from $d = 1$ to $d = 2$, the addition in the model of a new dimension for the DR subspace does not increase significantly the second largest eigen value. The observed value for λ_2^2 is close to λ_2^1 and is smaller than $1 - 2/(p + 1)$. Thus the procedure EIV based on (9) estimates here $d = 1$. The procedure based on AIC gives also $d = 1$ while the procedure based on BIC is not helpful here since it returns $d = 7$, thus does not indicates the existence of a DR subspace of small dimension. The fact that a single linear combination of the predictors carries all the information that X has about Y is a common conclusion in many works studying this dataset. Next we consider the DR subspace for $d = 1$.

We report in Table the estimated coefficients of Γ. These results suggest that the $6th$ and the $8th$ variables, namely the square root of the percent of mobility and the pupil-teacher ratio, are

	AFDC	BthPts	HS	Poverty	Minority	Mobility	Attend	PT.ratio
Non Stand.	-0.50	-1.60	1.72	3.39	-0.74	-0.22	-12.96	-0.04
Stand	-0.92	-1.15	1.62	1.90	-1.55	-0.18	1.03	-0.06

Table 3: Estimated coefficients of Γ. The first line gives the estimated coefficients returned by the method, the second line gives these coefficients standardized by the standard deviations of the original predictors.

maybe here negligible relatively to the other variables. A simple way to judge the validity of the estimated DR subspace is to verify visually if there exists an underlying function linking the single linear combination $X_0 = \Gamma^t X$ and Y. The scatterplot of the four responses and the estimated X_0 of the predictors is given in Figure 5. It suggests that, similarly to the results described in Yin and Bura [34], the responses variables could be described by monotonous quadratic or linear functions of X_0. On this scatterplot notice that the link between the response variables corresponding to the sixth grade reading comprehension tests and X_0 is more evident than the links between the variables corresponding to the fourth grade reading comprehension tests and X_0.

6 Conclusion

We have presented in this paper a model-based dimension reduction method. The model assumes that the whole joint distribution of (X, Y) is a finite mixture of distributions parametrized such that the matrix Γ is estimable. The model can handle multivariate response variables and is able to recover the DR subspace in the case of a regression symmetric relationship. A canonical choice is to use Gaussian distributions for the components of the mixture. We have presented for this case a procedure to estimate the parameters of the model, that involves an EM algorithm. In a simulation study the proposed method appeared to outperform existing ones for some designs and, globally, to performs at least equally to them.

We have also addressed in this paper the problem of the choice of the dimension d of the DR subspace. Following existing works for dimension reduction we have considered the use of the classical information criteria AIC and BIC. We have also proposed a simple procedure based on

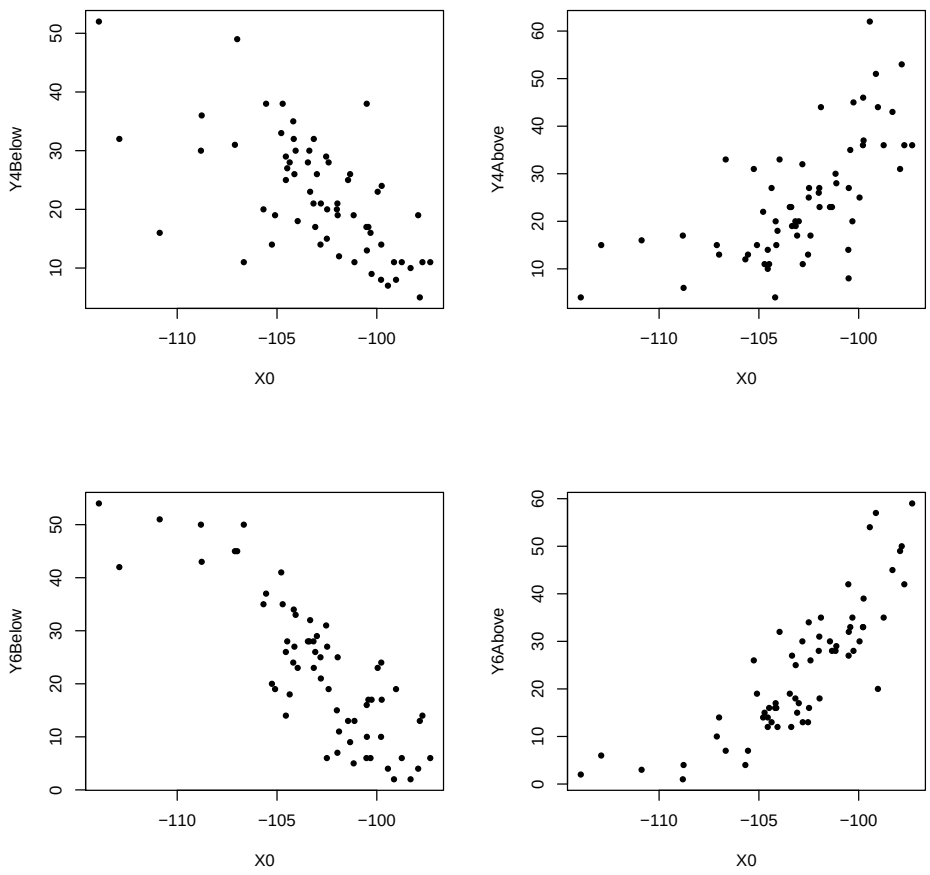

Figure 5: Minneapolis schools dataset. Scatterplot matrix of the four responses and the estimated single linear combination X_0 of the predictors.

the eigen values returned by the algorithm.

A future direction of work could concerns the extension of the model to handle non-continuous response variables. As noticed in Section 2 it is possible using appropriate functions $h_m(\cdot)$, $m = 1, \ldots, M$. Such an extension should be straightforward to consider binary regression or, more generally, the classification problem.

References

[1] Adragni, K. and Raim, A. (2014). ldr: An R Software Package for Likelihood-Based Sufficient Dimension Reduction. *Journal of Statistical Software*, **61**, 1–21.

[2] Aragon, Y. (1997). A Gauss implementation of multivariate Sliced Inverse Regression. *Computational Statistics*, **12**, 355–372.

[3] Bernard-Michel, C., Gardes, L. and Girard, S. (2009). Gaussian regularized sliced inverse regression. *Statistics and Computing*, **19**, 85–98.

[4] Bernard-Michel, C., Douté, S., 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.

[5] Bura, E. and Cook, R. D. (2001). Estimating the structural dimension of regressions via parametric inverse regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 393–410.

[6] Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983–992.

[7] Cook, R.D. (1998). *Regression graphics. Ideas for studying regressions through graphics.* Wiley Series in Probability and Statistics, New York

[8] Cook, R.D. (2007). Fisher lecture: Dimension reduction in regression. *Statistical Science*, **22(1)**, 1–26.

[9] Cook, R.D. and Forzani, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197 – 208.

[10] Cook, R.D. and Weisberg, S. (1991). Discussion of "Sliced Inverse Regression for dimension reduction". *Journal of the American Statistical Association*, **86**, 328–332.

[11] Coudret, R., Girard, S. and Saracco, J. (2014). A new sliced inverse regression method for multivariate response. *Computational Statistics and Data Analysis*, **77**, 285 – 299.

[12] Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, **39**, 1–38.

[13] Gannoun, A. and Saracco, J. (2003). An Asymptotic Theory for SIR_α Method. *Statistica Sinica*, **13**, 297–310.

[14] Hartigan, J.A. (1975). *Clustering algorithms.* John Wiley & Sons, New-York.

[15] Hsing, T. (1999). Nearest neighbor inverse regression. *The Annals of Statistics*, **27**, 697–731.

[16] Hsing, T. and Carroll, R.J. (1992). An asymptotic theory for Sliced Inverse Regression. *The Annals of Statistics*, **20**, 1040–1061.

[17] Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, **102**, 997–1008.

[18] Li, B., Zha, H. and Chiaromonte, F. (2005). Contour regression: A general approach to dimension reduction. *The Annals of Statistics*, **33**, 1580–1616.

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

[20] 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.

[21] Li, K.C., Aragon, Y., Shedden, K. and Thomas-Agnan, C. (2003). Dimension reduction for multivariate response data. *Journal of the American Statistical Association*, **98**, 99–109.

[22] Nadaraya, E.A. (1964). On estimating regression. *Theory of Probability and its Application*, **9(1)**, 141–142.

[23] Reich, B.J., Bondell, H.D. and Li, L. (2011). Sufficient dimension reduction via Bayesian mixture modeling. *Biometrics*, **67(3)**, 886–895.

[24] Saracco, J. (1997). An asymptotic theory for sliced inverse regression. *Communications in Statistics, Theory and Methods*, **26**, 2141–2171.

[25] Saracco, J. (2005). Asymptotics for pooled marginal slicing estimator based on SIR_α approach. *Journal of Multivariate Analysis*, **96(1)**, 117–135.

[26] Scrucca, L. (2011). Model-based SIR for dimension reduction. *Computational Statistics and Data Analysis*, **55**, 3010–3026.

[27] Setodji, C. and Cook, R. (2004). K-means inverse regression. *Technometrics*, **46**, 421–429.

[28] Szretter, M.E. and Yohai, V.J. (2009). The sliced inverse regression algorithm as a maximum likelihood procedure. *Journal of Statistical Planning and Inference*, **139**, 3570–3578.

[29] Wang, H. And Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, **103**, 811–821.

[30] Watson, G.S. (1964). Smooth regression analysis. *Sankhya Series A*, **26(15)**, 175–184.

[31] Weisberg, W. (2002). Dimension reduction regression in R. *Journal of Statistical Software*, **7**, 1–22.

[32] Xia, Y., Tong, H., Li, W.K. and Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, **64(3)**, 363–410.

[33] Ye, Z. and Weiss, R.E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, **98**, 968–979.

[34] Yin, X. and Bura, E. (2006). Moment-based dimension reduction for multivariate response regression. *Journal of Statistical Planning and Inference*, **136**, 3675–3688.

7 Appendix - Proof of Theorem 1

We first give a result on the eigenvalue decomposition of a product of two symmetric matrices that will be useful for the proof of Theorem 1.

Lemma 1. *Let A and B be $p \times p$ symmetric matrices and assume that A is regular.*

i) The matrix AB is diagonalizable with non-negative real eigenvalues $\lambda_1, \ldots, \lambda_p$.

ii) Let P be the orthonormal matrix whose columns are the eigenvectors of the symmetric matrix $A^{1/2}BA^{1/2}$. One has $Q^{-1}ABQ = L$ where $Q = A^{1/2}P$ and $L = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$.

iii) For each $d \leq p$, denote by U a $p \times d$ matrix whose columns are d different columns of Q. If $\max\{\lambda_1, \ldots, \lambda_p\} < 1$ the matrix $V = A^{-1} - A^{-1}ULU^tA^{-1}$ is a definite positive matrix.

Proof − The proof is based on the fact that AB is similar to the symmetric matrix $A^{1/2}BA^{1/2}$ since

$$A^{1/2}\left(A^{1/2}BA^{1/2}\right)A^{-1/2} = AB.$$

i) The proof is straightforward.

ii) Since $A^{1/2}BA^{1/2}$ is a symmetric matrix, there exist an orthonormal matrix P such that $PLP^t = A^{1/2}BA^{1/2}$. It is then straightforward to check that $Q^{-1}ABQ = L$ where $Q^{-1} = P^tA^{-1/2}$.

iii) Without loss of generality, assume that the columns of D are the eigenvectors associated to the d first eigenvalues of AB. Since $Q^tA^{-1}Q = I_p$, it is easy to check that $U^tA^{-1}Q = [I_d, 0_{d\times(p-d)}]$ where $0_{p\times q}$ is the zero $p \times q$ matrix. Using the fact that $Q^tA^{-1}Q = I_p$, the matrix Q^tVQ is then a diagonal matrix with diagonal given by $(1-\lambda_1, \ldots, 1-\lambda_d, 1, \ldots, 1)$. Since Q is regular and $\max\{\lambda_1, \ldots, \lambda_p\} < 1$ the conclusion is straightforward. ■

Before proving Theorem 1, let us recall the iterative procedure of the EM algorithm. Let $\Theta^{(s)}$, $\Gamma^{(s)}$ and $\{(\pi_m^{(s)}, \Theta_m^{(s)}),\ m = 1, \ldots, M\}$ be the estimated parameters of the model obtained at iteration s of the algorithm and let, for all $m = 1, \ldots, M$ and $i = 1, \ldots, n$,

$$z_{i,m}^{(s)} = \frac{\pi_m^{(s)} g_m((\Gamma^{(s)})^t x_i | \Gamma^{(s)}, \Theta_m^{(s)}) h_m(y_i | \Gamma^{(s)}, \Theta_m^{(s)})}{\sum_{j=1}^{M} \pi_j^{(s)} g_j((\Gamma^{(s)})^t x_i | \Gamma^{(s)}, \Theta_j^{(s)}) h_j(y_i | \Gamma^{(s)}, \Theta_j^{(s)})}.$$

In what follows we assume that, at each iteration of the algorithm, there is no empty component. At iteration $s+1$, the new estimated parameters $\Theta^{(s+1)}$, $\Gamma^{(s+1)}$ and $\{(\pi_m^{(s+1)}, \Theta_m^{(s+1)}),\ m = 1, \ldots, M\}$ are the values maximizing with respect to Θ, Γ and $\{(\pi_m, \Theta_m),\ m = 1, \ldots, M\}$ the expectation of the completed log-likelihood function

$$\sum_{i=1}^{n} \sum_{m=1}^{M} z_{i,m}^{(s)} \log \left[\pi_m g(x_i|\Theta) g_m(\Gamma^t x_i|\Gamma, \Theta_m) h_m(y_i|\Gamma, \Theta_m) \right]. \tag{10}$$

This procedure is iterated until convergence of the algorithm. This convergence is assessed by looking at the difference of contiguous estimates of the maximized log-likelihood.

The next result provides the expressions of the maximum likelihood estimators of the parameters of model (3). The following notations are required: for all $i = 1, \ldots, n$ and $m = 1, \ldots, M$, let $z_i = (z_{i,1}, \ldots, z_{i,M-1})^t \in \mathbb{R}^{M-1}$ and let D be the $p \times (M-1)$ matrix defined by:

$$D = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(z_i - \bar{z})^t \text{ with } \bar{z} = \frac{1}{n} \sum_{i=1}^{n} z_i.$$

We introduce the $(M-1) \times (M-1)$ matrix $F = \text{diag}(\bar{z}) - \bar{z}\bar{z}^t$. Note that F is regular with

$$F^{-1} = \text{diag}(1/\bar{z}) + \left(\frac{1}{n} \sum_{i=1}^{n} z_{i,M} \right)^{-1} \Omega, \tag{11}$$

where Ω is the $(M-1) \times (M-1)$ matrix whose entries are all equal to one. Denote L the $d \times d$ diagonal matrix of the d largest eigenvalues of the matrix $\hat{\Sigma}_n^{-1} D F^{-1} D^t$. According to Lemma 1, these eigenvalues are non-negative real values. Let also U, a $p \times d$ matrix corresponding to the first d columns of the matrix $\Sigma^{-1/2} A$ where A is the orthonormal matrix whose columns are the eigenvectors of the matrix $\Sigma^{-1/2} D F^{-1} D^t \Sigma^{-1/2}$. The columns of this matrix U are eigenvectors of the matrix $\hat{\Sigma}_n^{-1} D F^{-1} D^t$.

Proposition 1. *Under model (3), the maximum likelihood estimators are given by:*

i) $\hat{V} = \hat{\Sigma}_n - \hat{\Sigma}_n U L U^t \hat{\Sigma}_n$,

ii) $\hat{\Gamma} = U(U^t \hat{V} U)^{-1/2}$,

iii) the vectors $\hat{\beta}_1, \ldots, \hat{\beta}_{M-1}$ *that are the columns of the matrix* $\hat{\Gamma}^t D F^{-1}$,

iv) $\hat{\xi} = \bar{x} - \hat{V}\hat{\Gamma}\left(F^{-1}D^t\hat{\Gamma}\right)^t \bar{z}$,

v) $(\hat{\pi}_1, \ldots, \hat{\pi}_{M-1})^t = \bar{z}$,

vi) $\hat{\alpha}_m$ *and* $\hat{W}_m$ *that are the weighted maximum likelihood estimators maximizing:*

$$\sum_{i=1}^{n} z_{i,m}\left(-\frac{1}{2}\log|W_m| + \log h(W_m^{-1/2}(x-\alpha_m))\right).$$

Proof − Hereafter B denotes the $d \times (M-1)$ matrix with columns β_1,...,$\beta_{(M-1)}$. We also use the notations $\tilde{D} = D + \bar{x}\bar{z}^t$ and $\tilde{F} = F + \bar{z}\bar{z}^t$.

The values $\hat{\xi}$, $\hat{V}$, $\hat{B}$ and $\hat{\Gamma}$ maximizing the completed log-likelihood function (10) are the values minimizing the function

$$\begin{aligned} G(\xi, V, B, \Gamma) &= \ln|V| + \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^t V^{-1}(x_i-\bar{x}) + (\bar{x}-\xi)^t V^{-1}(\bar{x}-\xi) \\ &\quad -2Tr(\tilde{D}^t\Gamma B) + 2\xi^t\Gamma B\bar{z} + Tr(\tilde{F}B^t\Gamma^t V\Gamma B). \end{aligned}$$

Annulling the gradients of G leads to the equations

$$-\hat{V}^{-1}\bar{x} + \hat{V}^{-1}\hat{\xi} + \hat{\Gamma}\hat{B}\bar{z} = 0, \tag{12}$$

$$-\tilde{D}\hat{B}^t + \hat{\xi}\bar{z}^t\hat{B}^t + \hat{V}\hat{\Gamma}\hat{B}\tilde{F}\hat{B}^t = 0, \tag{13}$$

$$\hat{V}^{-1} - \hat{V}^{-1}\left[\hat{\Sigma}_n + (\bar{x}-\hat{\xi})(\bar{x}-\hat{\xi})^t\right]\hat{V}^{-1} + \hat{\Gamma}\hat{B}\tilde{F}\hat{B}^t\hat{\Gamma}^t = 0, \tag{14}$$

$$-\tilde{D}^t\hat{\Gamma} + \bar{z}\hat{\xi}^t\hat{\Gamma} + \tilde{F}\hat{B}^t\hat{\Gamma}^t\hat{V}\hat{\Gamma} = 0. \tag{15}$$

From (12) we get $\hat{\xi} = \bar{x} - \hat{V}\hat{\Gamma}\hat{B}\bar{z}$ and replacing in (13) and (14) leads to:

$$D\hat{B}^t = \hat{V}\hat{\Gamma}\hat{B}F\hat{B}^t, \tag{16}$$

$$\hat{V} = \hat{\Sigma}_n - \hat{V}\hat{\Gamma}\hat{B}F\hat{B}^t\hat{\Gamma}^t\hat{V}. \tag{17}$$

Furthermore, from (15), one has $\hat{B} = (\hat{\Gamma}^t\hat{V}\hat{\Gamma})^{-1}\hat{\Gamma}^t DF^{-1}$. Note that the function $G(\cdot)$ and the equations (12) to (15) are unchanged multiplying $\hat{\Gamma}$ by any $d \times d$ regular matrix. In what follows, we thus take $\hat{\Gamma}$ such that $\hat{\Gamma}^t\hat{V}\hat{\Gamma} = I_d$ proving iii). Replacing $\hat{B}$ in the expression of $\hat{\xi}$ leads to iv). If we multiply on the right (17) by $\hat{\Gamma}$, using the equality (16) and the constraint $\hat{\Gamma}^t\hat{V}\hat{\Gamma} = I_d$, we get

$$\hat{\Sigma}_n\hat{\Gamma} = \hat{V}\hat{\Gamma} + D\hat{B}^t \text{ and } \hat{V}\hat{\Gamma} = D\hat{B}^t(\hat{B}F\hat{B}^t)^{-1}.$$

Combining these equalities and replacing $\hat{B}$ by its expression entails that

$$\hat{\Gamma}T = \hat{\Sigma}_n^{-1} D F^{-1} D^t \hat{\Gamma}, \tag{18}$$

where we have introduced the $d \times d$ matrix T given by

$$T = \left[I_d + (\hat{B} F \hat{B}^t)^{-1} \right]^{-1} = \hat{B} F \hat{B}^t \left[I_d + \hat{B} F \hat{B}^t \right]^{-1}.$$

Notice that, multiplying (17) on the left by $\hat{\Gamma}^t$ and on the right by $\hat{\Gamma}$, we have $\hat{B} F \hat{B}^t + I_d = \hat{\Gamma}^t \hat{\Sigma}_n \hat{\Gamma}$ and thus

$$T = I_d - (\hat{\Gamma}^t \hat{\Sigma}_n \hat{\Gamma})^{-1}. \tag{19}$$

Since T is a symmetric matrix, there exists an orthonormal matrix Q such that $Q^t T Q$ is a diagonal matrix Δ. According to (18), $\hat{\Gamma} Q^t \Delta = \hat{\Sigma}_n^{-1} D F^{-1} D^t \hat{\Gamma} Q^t$ and thus, the d columns of $\hat{\Gamma}$ span an eigensubspace of the matrix $\hat{\Sigma}_n^{-1} D F^{-1} D^t$. Denoting by $\lambda_1, \ldots, \lambda_d$ the diagonal of Δ, we want to show that the minimun of $G(\xi, V, B, \Gamma)$ is given by

$$G(\hat{\xi}, \hat{V}, \hat{B}, \hat{\Gamma}) = \ln |\hat{\Sigma}_n| + \sum_{k=1}^{d} \ln(1 - \lambda_k) + p. \tag{20}$$

Note that, from (19),

$$\Delta = I_d - \left(Q \hat{\Gamma}^t \hat{\Sigma}_n \hat{\Gamma} Q^{-1} \right)^{-1}, \tag{21}$$

and thus, since $\hat{\Sigma}_n$ is definite positive, $\lambda_1, \ldots, \lambda_d$ are smaller than 1. A consequence of equality (20) is that $\hat{\Gamma}$ span the eigen subspace associated to the d largest eigenvalues of the matrix $\hat{\Sigma}_n^{-1} D F^{-1} D^t$. First, from (17), $\ln |\widehat{V}| = \ln |\hat{\Sigma}_n| - \ln |I_p + \hat{\Gamma}(\hat{\Lambda} + \hat{B} F \hat{B}^t) \hat{\Gamma}^t \hat{V}|$. Using Sylvester's identity we get

$$|I_p + \hat{\Gamma} \hat{B} F \hat{B}^t \hat{\Gamma}^t \hat{V}|^{-1} = |(\hat{B} F \hat{B}^t)^{-1} T| = |I_d - T|.$$

Thus, using (21), entails

$$\ln |\hat{V}| = \ln |\hat{\Sigma}_n| + \ln |I_d - \Delta| = \ln |\hat{\Sigma}_n| + \sum_{k=1}^{d} \ln(1 - \lambda_k). \tag{22}$$

Furthermore, using (17)

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^t \hat{V}^{-1} (x_i - \bar{x}) = Tr(\hat{\Sigma}_n \hat{V}^{-1}) = p + Tr(\hat{\Gamma}^t D F^{-1} D^t \hat{\Gamma}), \tag{23}$$

and, using iv) and the constraint on $\hat{\Gamma}$,

$$(\bar{x} - \hat{\xi})^t \hat{V}^{-1} (\bar{x} - \hat{\xi}) = \bar{z}^t \hat{B}^t \hat{B} \bar{z}. \tag{24}$$

Using iii),

$$Tr(\tilde{D}^t \hat{\Gamma} \hat{B}) = Tr(\hat{\Gamma}^t D F^{-1} \tilde{D}^t \hat{\Gamma}) = Tr(\hat{\Gamma}^t D F^{-1} D^t \hat{\Gamma}) + \bar{x}^t \hat{\Gamma} \hat{B} \bar{z}. \tag{25}$$

Finally, using iii), iv) and the constraint on $\hat{\Gamma}$ leads to

$$\hat{\xi}^t \hat{\Gamma} \hat{B} \bar{z} = \bar{x}^t \hat{\Gamma} \hat{B} \bar{z} - \bar{z}^t \hat{B}^t \hat{B} \bar{z} \tag{26}$$

and

$$Tr((F + \bar{z}\bar{z}^t) \hat{B}^t \hat{\Gamma}^t \hat{V} \hat{\Gamma} \hat{B}) = Tr(\hat{B} F \hat{B}^t) + Tr(\bar{z}^t \hat{B}^t \hat{B} \bar{z}) = Tr(\hat{\Gamma}^t D F^{-1} D^t \hat{\Gamma}) + \bar{z}^t \hat{B}^t \hat{B} \bar{z}. \tag{27}$$

Collecting (22) to (27) yield to equation (20). We are now interested in the proof of i) and ii). Let $V_0 = \hat{\Sigma}_n - \hat{\Sigma}_n U L U^t \hat{\Sigma}_n$ and take $\hat{\Gamma} = U(U^t V_0 U)^{-1/2}$. First, from Lemma 1, since all the eigenvalues of $\Sigma^{-1} D F^{-1} D^t$ are smaller than 1, the matrix V_0 is symmetric and definite positive. Next, using again Lemma 1 and equation (19), the matrix T is diagonal and thus $\hat{\Gamma}$ satisfies (18) with $T = L$ proving ii). Finally, it is easy to see that V_0 can be expressed as $V_0 = \hat{\Sigma}_n - \hat{\Sigma}_n \hat{\Gamma} L \hat{\Gamma}^t \hat{\Sigma}_n$ and simple calculations give that

$$V_0 + V_0 \hat{\Gamma} \hat{B} F \hat{B}^t \hat{\Gamma}^t V_0 = \hat{\Sigma}_n.$$

Hence, V_0 verifies (17) and i) is proved. The end of the proof is straightforward. ■

Remarks –

1) The maximum likelihood estimator of Γ is normalized in order to have $\hat{\Gamma}^t \hat{V} \hat{\Gamma} = I_d$.

2) If we consider the parcimonious model obtain by taking $V = \sigma^2 I_p$ with $\sigma^2 > 0$ an unknown parameter, the previous lemma remains true (with a very similar proof) by replacing $\hat{V}$ by $\hat{\sigma}^2 I_p$ where

$$\hat{\sigma}^2 = \frac{1}{n[p + Tr(D F^{-1} D^t U U^t)]} \sum_{i=1}^{n} (x_i - \bar{x})^t (x_i - \bar{x}).$$

3) If we take for $h(\cdot)$ the q-dimensional pdf of a standard Gaussian distribution, the maximum likelihood estimators in vi) are given for $m = 1, \ldots, M$ by:

$$\hat{\alpha}_m = \sum_{i=1}^{n} z_{i,m} y_i \Big/ \sum_{i=1}^{n} z_{i,m} \quad \text{and } \hat{W}_m = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{M} z_{i,m} (y_i - \hat{\alpha}_m)(y_i - \hat{\alpha}_m)^t.$$

Assuming that $W_m = v^2 I_q$, the maximum likelihood estimator of W_m is $\hat{v}^2 I_q$ with

$$\hat{v}^2 = \frac{1}{nq} \sum_{i=1}^{n} \sum_{m=1}^{M} z_{i,m} (y_i - \hat{\alpha}_m)^t (y_i - \hat{\alpha}_m).$$

Proof of Theorem 1 – It suffices to show that

$$DF^{-1}D^t = \sum_{m=1}^{M} \hat{\pi}_m (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^t.$$

First, using the point v) of Lemma 1, the j-th column of D is given by

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(z_{i,j} - \hat{\pi}_j) = \hat{\pi}_j (\bar{x}_j - \bar{x}).$$

Hence, from (11),

$$DF^{-1}D^t = \sum_{m=1}^{M-1} \hat{\pi}_m (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^t + \frac{1}{\hat{\pi}_M} D\Omega D^t.$$

Remark that all the columns of $D\Omega$ are given by

$$\sum_{m=1}^{M-1} \hat{\pi}_m (\bar{x}_m - \bar{x}) = -\hat{\pi}_M (\bar{x}_M - \bar{x}),$$

since

$$\sum_{m=1}^{M} \hat{\pi}_m (\bar{x}_m - \bar{x}) = \bar{x} - \bar{x} = 0.$$

Taking account of $\Omega^2 = (M-1)\Omega$,

$$\frac{1}{\hat{\pi}_M} D\Omega D^t = \frac{1}{(M-1)\hat{\pi}_M} (D\Omega)(D\Omega)^t = \hat{\pi}_M (\bar{x}_m - \bar{x})(\bar{x}_m - \bar{x})^t,$$

which concludes the proof. ■