A Robust Proposal for Sliced Inverse Regression

V. Yohai[1] and M.E. Szretter Noste[2]

[1] Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Pabellón I, C1428EHA Buenos Aires, Argentina

[2] Instituto de Cálculo, Facultad de Ciencias Exactas y Naturales, Ciudad Universitaria, Pabellón II, C1428EHA Buenos Aires, Argentina

Keywords: Robustness, SIR, Dimension reduction.

Abstract

In this work we study the procedure of dimension reduction for multivariate observations known as Sliced Inverse Regression (SIR) presented by K. C. Li (1991). We prove that the algorithm developed by Li (1991) to solve the problem of sliced inverse regression provides the same results as those obtained by the maximum likelihood method of estimating the subspace that contains the means of the groups induced by the slicing of the observations, under the assumption of normality and equal covariance matrix.

The model presented by Li is to estimate, from a sample $(\mathbf{x_i}', y_i)'$, $1 \leq i \leq N$, a non parametric relation between $\mathbf{x}$ and y, where p, the dimension of $\mathbf{x}$ is big. Of course, this is not possible except in the case that N is big enough to overcome the curse of dimensionality. Li proposes an alternative way to avoid this problem, the following non parametric model where y depends on $\mathbf{x}$ only through a reduced number K of lineal combinations

$$y = f\left(\beta_1'\mathbf{x}, \ldots, \beta_K'\mathbf{x},\varepsilon\right) \tag{1}$$

where y is the response variable, $\mathbf{x}$ is the p-dimensional vector of covariables, ε is the error, which is independent from $\mathbf{x}$, β_i are unknown vectors in $\mathbf{R}^p$ and f is an arbitrary function, $f : \mathbf{R}^{K+1} \to \mathbf{R}$. Li's proposal is based on the idea of inverse regression: to put a model on $E(\mathbf{x} \mid y)$ and focus on the estimation of the β_i's. Li proves that under a general condition this curve, once centered, falls in a subspace of dimension K in $\mathbf{R}^p$ linearly related to de the subspace of interest, the edr (effective dimension reduction) space, which is the one generated by $\beta_1, ..., \beta_K$.

This suggests an alternative method to estimate the this subspace SIR model, assuming that the observations $\mathbf{x}_i$, which are classified in groups (slices) according to the value of variable y, have a multivariate normal distribution normal with means belonging to a K-dimensional affine variety in $\mathbf{R}^p$, and the same covariance matrix Σ, that is

$$\mathbf{x}_{hj} \sim N_p\left(\alpha_h, \Sigma\right), \ 1 \leq j \leq n_h, 1 \leq h \leq H, \quad \alpha_h \in V + \mathbf{a}, \dim(V) = K, V \subset \mathbf{R}^p \tag{2}$$

Then we estimate α_h and Σ by maximun likelihood method, and estimate the edr subspace from these. We obtain the following MLE:

$$\hat{\Sigma} = W + B^{1/2}C \begin{bmatrix} I_{(p-K)} & 0_{(p-K)\times K} \\ 0_{K\times(p-K)} & 0_{K\times K} \end{bmatrix} C'B^{1/2} \tag{3}$$

where

$$B = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)\left(\overline{\mathbf{x}}_{i\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right)' \tag{4}$$

$$W = \frac{1}{N}\sum_{i=1}^{H}\sum_{j=1}^{n_i}\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)\left(\mathbf{x}_{ij} - \overline{\mathbf{x}}_{i\bullet}\right)', \tag{5}$$

and C is the orthogonal matrix obtained by the spectral decomposition of

$$B^{-1/2}WB^{-1/2} = C\Omega C' \tag{6}$$

where Ω is the diagonal matrix that contains the eigenvalues, ordered in a decreasing way. The MLE of α_h are

$$\widehat{\alpha}_h = \widehat{\Sigma}^{1/2}DD'\widehat{\Sigma}^{-1/2}\left(\overline{\mathbf{x}}_{h\bullet} - \overline{\mathbf{x}}_{\bullet\bullet}\right) + \overline{\mathbf{x}}_{\bullet\bullet} \tag{7}$$

where $D = [\mathbf{t}_1 \cdots \mathbf{t}_K] \in \mathbf{R}^{p\times K}$ with $\{\mathbf{t}_1,\ldots,\mathbf{t}_K\}$ the orthogonal eigenvectors of $\widehat{\Sigma}^{-1/2}B\widehat{\Sigma}^{-1/2}$ associated to the K largest eigenvalues. Finally, the estimates for the edr directions are

$$\widehat{\beta}_k = \widehat{\Sigma}_{\mathbf{xx}}^{-1}\widehat{\Sigma}^{1/2}\mathbf{t}_k, \tag{8}$$

$k = 1,\ldots,K$, where $\widehat{\Sigma}_{\mathbf{xx}}$ is the sample covariance matrix of $\mathbf{x}$. We also show that the subspace that contains the inverse regression curve estimated by Li's algorithm from a sample, and the one that contains the means of every slice of the $\mathbf{x}_i$ estimated by ML are the same one.

The usefulness of this approach lies on the possibility of viewing the estimators within a maximum likelihood strategy. This enables to apply to the obtained estimators the well known properties of this general estimation method. Also, it allows to search for an alternative way of finding a robust estimator, different from those proposed so far (Gather, U., Hilker, T. y Becker, C. (2001, 2002)). This procedure that finds robust estimators for a model that allows maximum likelihood estimation was employed by García Ben, Martínez and Yohai (2004) to propose robust and efficient estimates for multivariate lineal models. We define the $\tau-$ estimator for the SIR model by

$$\left(\widetilde{\alpha}_1,\ldots,\widetilde{\alpha}_H,\widetilde{\Sigma}\right) = \arg\min_{\Sigma>0;\alpha_1,\ldots,\alpha_H\in V+\mathbf{a}}|\Sigma| \tag{9}$$

subject to

$$\tau^2\left(d_{11}\left(\alpha_1,\Sigma\right),\ldots,d_{Hn_H}\left(\alpha_H,\Sigma\right)\right) = \tau_0^2 \tag{10}$$

$$\alpha_1,\ldots,\alpha_H \in V + \mathbf{a} \tag{11}$$

where τ is robust scale $\tau-$estimator, V is a subspace of dimension K, τ_0 is a constant selected to be consistent under normality, and d_{ij} are the Mahalanobis distances

$$d_{ij}^2\left(\alpha_i,\Sigma\right) = \left(\mathbf{x}_{ij} - \alpha_i\right)'\Sigma^{-1}\left(\mathbf{x}_{ij} - \alpha_i\right) \tag{12}$$

The resulting robust estimators will be presented, together with a Montecarlo simulation.

References

M. García Ben, E.J. Martínez and V.J. Yohai (2004). Robust and Efficient Estimates for Multivariate Lineal Models, unpublished article.

U. Gather, T. Hilker and C. Becker (2001). A Robustified Version of Sliced Inverse Regression. In: L.T. Fernholz, S. Morgenthaler and W. Stahel, editors, *Statistics in Genetics and in the Environmental Sciences, Proceedings of the Workshop on Statistical Methodology for the Sciences: Environmetrics and Genetics* held in Ascona from May 23 to 28, pp. 147–157.

U. Gather, T. Hilker and C. Becker (2002). A note on outlier sensitivity of Sliced Inverse Regression, *Statistics*, 13(4), 271–281.

K.C. Li (1991). Sliced Inverse Regression for Dimension Reduction, *Journal of the American Statistical Association*, 86, 316–327.