ARTICLE IN PRESS
Computational Statistics and Data Analysis ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Student Sliced Inverse Regression

Alessandro Chiancone [a,b,c,*], Florence Forbes [a], Stéphane Girard [a]

[a] Inria Grenoble Rhône-Alpes & LJK, team Mistis, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France
[b] GIPSA-Lab, Grenoble INP, Saint Martin d'Hères, France
[c] Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, A-8010 Graz, Austria

ARTICLE INFO

Article history:
Received 24 February 2016
Received in revised form 8 August 2016
Accepted 9 August 2016
Available online xxxx

Keywords:
Dimension reduction
Inverse regression
Outliers
Robust estimation
Generalized Student distribution

ABSTRACT

Sliced Inverse Regression (SIR) has been extensively used to reduce the dimension of the predictor space before performing regression. SIR is originally a model free method but it has been shown to actually correspond to the maximum likelihood of an inverse regression model with Gaussian errors. This intrinsic Gaussianity of standard SIR may explain its high sensitivity to outliers as observed in a number of studies. To improve robustness, the inverse regression formulation of SIR is therefore extended to non-Gaussian errors with heavy-tailed distributions. Considering Student distributed errors it is shown that the inverse regression remains tractable via an Expectation–Maximization (EM) algorithm. The algorithm is outlined and tested in the presence of outliers, both in simulated and real data, showing improved results in comparison to a number of other existing approaches.

1. Introduction

Let us consider a regression setting where the goal is to estimate the relationship between a univariate response variable Y and a predictor $\mathbf{X}$. When the dimension p of the predictor space is 1 or 2, a simple 2D or 3D plot can visually reveal the relationship and can be useful to determine the regression strategy to be used. If p becomes large such an approach is not feasible. A possibility to overcome problems arising in the context of regression is to make the assumption that the response variable does not depend on the whole predictor space but just on a projection of $\mathbf{X}$ onto a subspace of smaller dimension. Such a dimensionality reduction leads to the concept of sufficient dimension reduction and to that of central subspace (Cook, 1996). The central subspace is the intersection of all dimension-reduction subspaces (d.r.s.). A subspace S is a d.r.s. if Y is independent of $\mathbf{X}$ given $\mathbf{P}_S\mathbf{X}$, where $\mathbf{P}_S$ is the orthogonal projection onto S. In other words, all the information carried by the predictors $\mathbf{X}$ on Y can be compressed in $\mathbf{P}_S\mathbf{X}$. It has been shown under weak assumptions that the intersection of all d.r.s., and therefore the central subspace, is itself a d.r.s. (Yin et al., 2008). It is of particular interest to develop methods to estimate the central subspace as once it is identified, the regression problem can be solved equivalently using the lower-dimensional representation $\mathbf{P}_S\mathbf{X}$ of $\mathbf{X}$ in the subspace.

Among methods that lead to an estimation of the central subspace, Sliced Inverse Regression (SIR) (Li, 1991) is one of the most popular. SIR is a semiparametric method assuming that the link function depends on d linear combinations of the predictors and a random error independent of $\mathbf{X}$: $Y = f(\beta_1^T\mathbf{X}, \dots, \beta_d^T\mathbf{X}, \epsilon)$. When this model holds, the projection of $\mathbf{X}$ onto the space spanned by the vectors $\{\beta_i,\ i = 1, \dots, d\}$ captures all the information about Y. In addition, Li (1991) shows that a basis of this space can be recovered using an inverse regression strategy provided that the so called *linearity condition* holds. It has been shown that the *linearity condition* is satisfied as soon as $\mathbf{X}$ is elliptically distributed. Moreover, this condition

* Corresponding author at: Inria Grenoble Rhône-Alpes & LJK, team Mistis, 655, av. de l'Europe, Montbonnot, 38334 Saint-Ismier cedex, France.
E-mail address: al.chiancone@gmail.com (A. Chiancone).

http://dx.doi.org/10.1016/j.csda.2016.08.004
0167-9473/

approximately holds in high-dimensional datasets, see Hall and Li (1993). However, solutions have been proposed to deal with non elliptical distributed predictors and to overcome the *linearity condition* limitation (Fukumizu et al., 2004; Li and Dong, 2009; Fukumizu et al., 2009).

The inverse regression approach to dimensionality reduction gained then rapid attention (Cook and Weisberg, 1991) and was generalized in Cook (2007) which shows the link between the axes spanning the central subspace and an inverse regression problem with Gaussian distributed errors. More specifically, in Bernard-Michel et al. (2009) and Cook (2007), it appears that, for a Gaussian error term and under appropriate conditions, the SIR estimator can be recovered as the maximum likelihood estimator of the parameters of an inverse regression model. In other words, although SIR is originally a model free method, the standard SIR estimates are shown to correspond to maximum likelihood estimators for a Gaussian inverse regression model. It is then not surprising that SIR has been observed, *e.g.* in Bura and Cook (2001), to be at best under normality and that its performance may degrade otherwise. Indeed, the Gaussian distribution is known to have tails too light to properly accommodate extreme values. In particular, Sheather and McKean (1997) observes that SIR was highly sensitive to outliers, with additional studies, evidence and analysis given in Gather et al. (2002). To downweight this sensitivity, robust versions of SIR have been proposed, mainly starting from the standard *model free* estimators and trying to make them more resistant to outliers. Typically, in Gather et al. (2001) classical estimators are replaced by high breakdown robust estimators and, recently in Dong et al. (2015) two approaches are built: a weighted version of SIR and a solution based on the intra slice multivariate median estimator.

As an alternative, we propose to rather exploit the inverse regression formulation of SIR (Bernard-Michel et al., 2009; Cook, 2007). A new error term modeled by a multivariate Student distribution (Dong et al., 2015) is introduced. Among the elliptically contoured distributions, the multivariate Student is a natural generalization of the multivariate Gaussian but its heavy tails can better accommodate outliers. The result in Proposition 6 of Cook (2007) is extended from Gaussian to Student errors showing that the inverse regression approach of SIR is still valid outside the Gaussian case, meaning that the central subspace can still be estimated by maximum likelihood estimation of the inverse regression parameters. It is then shown that the computation of the maximum likelihood estimators remains tractable in the Student case via an Expectation–Maximization (EM) algorithm which has a simple implementation and desirable properties.

The paper is organized as follows. In Section 2 general properties of the multivariate Student distribution and some of its variants are first recalled. The inverse regression model is introduced in Section 3 followed by the EM strategy to find the maximum likelihood estimator, the link with SIR and the resulting Student SIR algorithm. A simulation study is carried out in Section 4 and a real data application, showing the interest of this technique, is detailed in Section 5. The final section contains concluding remarks and perspectives. Proofs are postponed to the Appendix.

2. Multivariate generalized Student distributions

Multivariate Student, also called t-distributions, are useful when dealing with real-data because of their heavy tails. They are a robust alternative to the Gaussian distribution, which is known to be very sensitive to outliers. In contrast to the Gaussian case though, no closed-form solution exists for the maximum likelihood estimation of the parameters of the t-distribution. Tractability is, however, maintained both in the univariate and multivariate case, via the EM algorithm (McLachlan and Peel, 2000) and thanks to a useful representation of the t-distribution as a so-called *infinite mixture of scaled Gaussians* or *Gaussian scale mixture* (Andrews and Mallows, 1974). A Gaussian scale mixture distribution has a probability density function of the form

$$P(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\psi}) = \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\, f_U(u; \boldsymbol{\psi})\, \mathrm{d}u, \tag{1}$$

where $\mathcal{N}_p(\,.\,; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)$ denotes the density function of the p-dimensional Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}/u$ and f_U is the probability distribution of a univariate positive variable U referred to hereafter as the weight variable. When f_U is a Gamma distribution $\mathcal{G}(\nu/2, \nu/2)$ where ν denotes the degrees of freedom, expression (1) leads to the standard p-dimensional t-distribution denoted by $t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$ with parameters $\boldsymbol{\mu}$ (location vector), $\boldsymbol{\Sigma}$ ($p \times p$ positive definite scale matrix) and ν (positive degrees of freedom parameter). Its density is given by

$$\begin{aligned} t_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) &= \int_0^\infty \mathcal{N}_p(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\, \mathcal{G}(u; \nu/2, \nu/2)\, \mathrm{d}u \\ &= \frac{\Gamma((\nu+p)/2)}{|\boldsymbol{\Sigma}|^{1/2}\, \Gamma(\nu/2)\, (\pi\nu)^{p/2}}\, [1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\nu]^{-(\nu+p)/2}, \end{aligned} \tag{2}$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$ is the Mahalanobis distance between $\boldsymbol{x}$ and $\boldsymbol{\mu}$. The Gamma distribution has probability density function $\mathcal{G}(u; \alpha, \gamma) = u^{\alpha-1} \Gamma(\alpha)^{-1} \exp(-\gamma u) \gamma^\alpha$, where Γ denotes the Gamma function.

If $f_U(u; \boldsymbol{\psi})$ is set equal to a Gamma distribution $\mathcal{G}(\alpha, \gamma)$ without imposing $\alpha = \gamma$, (1) results in a multivariate Pearson type VII distribution (see *e.g.* Johnson et al., 1994 vol. 2 chap. 28) also referred to as the Arellano-Valle and Bolfarine's Generalized t distribution in Dong et al. (2015). This generalized version is the multivariate version of the t-distribution considered in

Please cite this article in press as: Chiancone, A., et al., Student Sliced Inverse Regression. Computational Statistics and Data Analysis (2016), http://dx.doi.org/10.1016/j.csda.2016.08.004