Computational Statistics and Data Analysis 125 (2018) 136–155

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Supervised dimension reduction for ordinal predictors☆

Liliana Forzani [a], Rodrigo García Arancibia [b,*], Pamela Llop [a], Diego Tomassi [a]

[a] Facultad de Ingeniería Química, Universidad Nacional del Litoral, Researchers of CONICET, Argentina
[b] Instituto de Economía Aplicada Litoral (FCE-UNL) and CONICET, Argentina

ARTICLE INFO

Article history:
Received 29 May 2017
Received in revised form 27 March 2018
Accepted 30 March 2018
Available online 12 April 2018

Keywords:
Expectation–maximization (EM)
Latent variables reduction subspace
SES index construction
Supervised classification
Variable selection

ABSTRACT

In applications involving ordinal predictors, common approaches to reduce dimensionality are either extensions of unsupervised techniques such as principal component analysis, or variable selection procedures that rely on modeling the regression function. A supervised dimension reduction method tailored to ordered categorical predictors is introduced which uses a model-based dimension reduction approach, inspired by extending sufficient dimension reductions to the context of latent Gaussian variables. The reduction is chosen without modeling the response as a function of the predictors and does not impose any distributional assumption on the response or on the response given the predictors. A likelihood-based estimator of the reduction is derived and an iterative expectation–maximization type algorithm is proposed to alleviate the computational load and thus make the method more practical. A regularized estimator, which simultaneously achieves variable selection and dimension reduction, is also presented. Performance of the proposed method is evaluated through simulations and a real data example for socioeconomic index construction, comparing favorably to widespread use techniques.

1. Introduction

Regression models with ordinal predictors are common in many applications. For instance, in economics and the social sciences, ordinal variables are used to predict phenomena like income distribution, poverty, consumption patterns, nutrition, fertility, healthcare decisions, and subjective well-being, among others (e.g. Bollen et al., 2001; Roy and Chaudhuri, 2009; Murasko, 2007; Kamakura and Mazzon, 2013; Mazzonna, 2014; Feeny et al., 2014). In marketing research, customer preferences are used to create automatic recommendation systems, as in the case of Netflix (e.g. Bobadilla et al., 2012; Roberts, 2014), where the ratings for unseen movies can be predicted using the user's previous ratings and information about the consumer preferences for the whole database.

In this context, when the number of predictors is large, it is of interest to reduce the dimensionality of the space by combining them into a few variables in order to get efficiency in the estimation as well as an understanding of the model. The commonly used dimension reduction techniques for ordinal variables are adaptations of standard principal component analysis (PCA) (Linting and van der Kooij, 2009; Kolenikov and Angeles, 2009). For example, in the case of the *Índice de Focalización de Pobreza* (a socioeconomic index commonly used in Latin America), the first normalized principal component is used to predict poverty status, even if this outcome variable was never used to estimate the scaling. It is clear, however, that ignoring the response when building such an index can lead to a loss of predictive power compared to the full set of predictors.

☆ Matlab codes for the algorithm and simulations are available at http://bit.ly/2xAWXTT.

* Corresponding author.
E-mail address: rgarcia@fce.unl.edu.ar (R. García Arancibia).

https://doi.org/10.1016/j.csda.2018.03.018
0167-9473/

A different approach to dimensionality reduction is to perform variable selection on the original set of predictors. A method adapted to ordinal predictors is proposed in Gertheiss and Tutz (2010). Despite the fact that this method uses information from the response to achieve variable selection, it performs simultaneously regression modeling by assuming a parametric model for the response as a function of predictors.

For regression and classification tasks it is widely accepted that supervised dimension reduction is a better alternative than PCA-like approaches. Sufficient dimension reduction (SDR), in particular, has gained interest in recent years as a principled methodology to achieve dimension reduction on the predictors $\mathbf{X} \in \mathbb{R}^p$ without losing information about the response Y. Formally, for the regression of $Y|\mathbf{X}$, SDR amounts to finding a transformation $R(\mathbf{X}) \in \mathbb{R}^d$, with $d \leq p$, such that the conditional distribution of $Y|\mathbf{X}$ is identical to that of $Y|R(\mathbf{X})$. Nevertheless, there is no need to assume a distribution for Y or for $Y|\mathbf{X}$. Thus, the obtained reductions can subsequently be used with any prediction rule. Moreover, when the reduced space has low dimension, it is feasible to plot the response versus the reduced variables. This can play an important role in facilitating model building and understanding (Cook, 1996, 1998).

Most of the methodology in SDR is based on the inverse regression of $\mathbf{X}$ on Y, which translates a p-dimensional problem of regressing $Y|\mathbf{X}$ into p (easier to model) one-dimensional regressions corresponding to $\mathbf{X}|Y$. Estimation in SDR was developed originally for continuous predictors and was based on the moments of the conditional distribution of $\mathbf{X}|Y$ (SIR, (Li, 1991); SAVE, (Cook and Weisberg, 1991); pHd, (Li, 1992); PIR (Bura and Cook, 2001); MAVE, (Xia et al., 2002; Li et al., 2005; Cook and Ni, 2005; Zhu and Zeng, 2006; Cook and Li, 2002); DR, (Li and Wang, 2007), see also Cook and Lee (1999), Cook and Yin (2001), Chiaromonte et al. (2002), Cook and Ni (2005), Yin et al. (2008) and Cook (1998), where much of its terminology was introduced). Later, Cook (2007) introduced the so-called *model-based* inverse regression of $\mathbf{X}|Y$ (see also Cook and Forzani, 2008, 2009). The main advantage of this approach is that provides an estimator of the sufficient reduction that contains all the information in $\mathbf{X}$ that is relevant to Y, allowing maximum likelihood estimators which are optimal in terms of efficiency and $\sqrt{n}$-consistent under mild conditions when the model holds.

Along the lines of the model-based SDR approach, the up to date methodology is for predictors belonging to a general exponential family of distributions (see Bura et al., 2016). Then, when attempting to apply SDR to ordinal predictors, a first approach could be to treat them as polytomic variables, ignoring their natural order. Then a multinomial distribution can be postulated over them, which can be treated as a member of the exponential family. However, ordered variables usually do not follow a multinomial distribution and the order information is lost when treating them as multinomial. There have been attempts to use dummy variables to deal with ordinal data, but this procedure has been shown to introduce spurious correlations (Kolenikov and Angeles, 2009). Another approach is to treat the ordered predictors as a discretization of some underlying continuous random variable. This technique is commonly used in the social sciences and is known as the *latent variable model*. In this context, the latent variables are usually modeled as normally or logistically distributed, obtaining the so-called ordered probit and logit models, respectively (Greene and Hensher, 2010; Long, 1997). While for each scientific phenomenon the latent variable can take a particular meaning (e.g., utility in economic choice problems, liability to diseases in genetics, or tolerance to a drug in toxicology), a general interpretation of a latent variable may be the *propensity* to observe a certain value j of an ordered categorical variable (Skrondal and Rabe-Hesketh, 2004). Regardless of their philosophical meaning and the criticisms about their real existence, latent variables are very useful for generating distributions for modeling, hence their widespread use.

In this paper, we develop a supervised dimension reduction method for ordinal predictors, based on the SDR for the regression of the response given the underlying normal latent variables. Under this context, we present a maximum likelihood estimator of the reduction and we propose an approximate expectation–maximization (EM) algorithm for its practical computation, which is close to recent developments in graphical models for ordinal data (Guo et al., 2015) and allows for computationally efficient estimation without losing accuracy.

The rest of this paper is organized as follows. In Section 2 we describe the inverse regression model for ordinal data and its dimensionality reduction. In Section 3 we derive the Maximum likelihood estimates of the reduction and we also present a variable selection method. Section 4 is dedicated to developing a permutation test for choosing the dimension for the reduction. Simulation results are presented in Section 5. Section 6 contains a socioeconomic application using the methodology developed in this paper to create a socioeconomic status (SES) index from ordinal predictors. Finally, a concluding discussion is given in Section 7. All proofs and other supporting material are given in the appendices.

2. Model

Let us consider the regression of a response $Y \in \mathbb{R}$ on a predictor $\mathbf{X} = (X_1, X_2, \ldots, X_p)^T$, where each $X_j, j = 1, \ldots, p$ is an ordered categorical variable, i.e., $X_j \in \{1, \ldots, G_j\}, j = 1, \ldots, p$. To state a dimension reduction of $\mathbf{X}$ inspired by the model-based SDR approach (see Cook, 2007), we should model the inverse regression of $\mathbf{X}$ on Y. However, as we stated in the introduction, the model-based SDR techniques deal with continuous predictors. Therefore, in order to frame our problem in that context, we will assume the existence of a p-dimensional vector of unobserved underlying continuous latent variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_p)^T$, with $E(\mathbf{Z}) = 0$, such that each observed X_j is a discretizing of Z_j as follows. There exists a set of thresholds $\Theta^{(j)} = \{\theta_0^{(j)}, \theta_1^{(j)}, \ldots, \theta_{G_j}^{(j)}\}$, that split the real line in disjoints intervals $-\infty = \theta_0^{(j)} < \theta_1^{(j)} < \cdots < \theta_{G_{j-1}}^{(j)} < \theta_{G_j}^{(j)} = +\infty$ and

$$X_j = \sum_{g=1}^{G_j} g\mathbb{I}(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)}), \tag{1}$$

where $\mathbb{I}(A)$ is the indicator function of the set A. Therefore, $X_j = g \Leftrightarrow Z_j \in [\theta_{g-1}^{(j)}, \theta_g^{(j)})$ and $P(X_j = g) = P(\theta_{g-1}^{(j)} \leq Z_j < \theta_g^{(j)})$.