THE MULTIVARIATE CLUSTER-WEIGHTED DISJOINT FACTOR ANALYZERS MODEL

Francesca Martella [1], Xiaoke Qin[2] and Wangshu Tu[2] and Sanjena Subedi[2]

[1] Department of Statistical Sciences, Sapienza University of Rome, (e-mail: francesca.martella@uniroma1.it)

[2] School of Mathematics and Statistics, Carleton University, (e-mail: XIAOKEQIN@cmail.carleton.ca, wangshu.tu@carleton.ca, Sanjeena.Dang@carleton.ca)

ABSTRACT: Cluster-weighted factor analyzers (CWFA) models are a flexible family of mixture models for fitting the joint distribution of a random vector constituted by a response variable and a set of explanatory variables. It is a useful tool especially when high-dimensionality and multicollinearity occurs. This paper extends CWFA models in two significant ways. Firstly, it allows to predict more than one response variable accounting for their potential interactions. Secondly, it identifies factors that relate to disjoint clusters of explanatory variables, simplifying their interpretatiblity. This leads to the multivariate cluster-weighted disjoint factor analyzers (MCWDFA) model. An alternating expectation-conditional maximization algorithm is used for parameter estimation. Application of the proposed approach to both simulated and real datasets is presented.

KEYWORDS: finite mixtures, factor regression model, disjoint factor analysis.

1 Introduction

Mixture models represent a powerful statistical tool for clustering observations which is an essential task in many fields, such as economics, engineering, and social sciences. In the context of media technology, Gershenfeld, 1997 proposed a particular family of Gaussian mixture models, called cluster-weighted models (CWMs), which has also been called saturated mixture regression models in Wedel, 2002. The context of interest is represented by data arising from a random vector $(\mathbf{X}, Y)'$, in which a functional dependence of Y on $\mathbf{X}$ is assumed for each mixture-component and the component-specific joint density of $(\mathbf{X}, Y)'$ is factorized into the product of the conditional density of $Y|\mathbf{X}$ and the marginal density of $\mathbf{X}$. Ingrassia *et al.*, 2012 reformulated the CWM in a statistical setting under the assumptions that both the component-specific conditional distributions of $Y|\mathbf{X}$ and the component-specific marginal

distributions of $\mathbf{X}$ are Gaussian. To allow the applicability of CWM in high dimensional $\mathbf{X}$-spaces or when multicollinearity occours, Subedi *et al.*, 2013 proposed the cluster-weighted factor analyzers (CWFA) model, which addressed the problem by assuming a latent structure for the explanatory variables in each mixture component. The aim of this paper is to propose a new model, called the multivariate cluster-weighted disjoint factor analyzers (MCWDFA) model, extending CWFA model in a two fold way. Firstly, it allows to predict more than one response variable accounting for their potential interactions. It leads to a more flexible model since it can capture the complexity and variability of real phenomena more accurately providing a more complete understanding of the underlying mechanisms of a case study. Secondly, it identifies factors that relate to disjoint clusters of explanatory variables which similarly predict the responses. In particular, following the idea of Martella *et al.*, 2008 and Vichi, 2017, we replace the factor loading matrix with the product of a binary row-stochastic matrix and a diagonal matrix in the factor analyzer structure. In this way, the explanatory variables that similarly predict the responses can be clustered into groups such that an explanatory variable loads only on one single factor, and thus, it is uniquely associated by a single factor only. This simplifies not only the interpretability of the resulting factors but also the interpretability of the (many) regression coefficients, especially when the explanatory variables matrices are in high-dimensional $\mathbf{X}$-spaces.

2 The cluster-weighted factor analyzers model

Briefly, the CWFA model (Subedi *et al.*, 2013) is a particular mixture model for fitting the joint distribution of a random vector composed of a response variable and a set of explanatory variables, where, within each Gaussian in the mixture, a single factor analysis regression (FAR) model (Basilevsky, 1981) is assumed. Let $y \in \mathbb{R}$ and $\mathbf{X} \in \mathbb{R}^p$ be a response variable and a vector of explanatory variables, respectively, realizations of the pair $(\mathbf{X}, Y)$. Specifically, the CWFA model postulates that:

$$Y = \beta_{0g} + \beta'_{1g}\mathbf{X} + e_g \qquad \text{with} \qquad \mathbf{X} = \mu_g + \Lambda_g \mathbf{F}_g + \varepsilon_g \tag{1}$$

with probability π_g ($g = 1, \ldots, G$). Terms μ_g represents the component-specific mean vectors of $\mathbf{X}$, Λ_g is a $p \times Q$ component-specific factor loadings matrix ($Q < p$), $\mathbf{F}_g$ is a Q-dimensional vector of component-specific factors, which are assumed to be i.i.d. draws from a Gaussian distribution $N(0, \mathbf{I}_Q)$ and $\mathbf{I}_Q$ denotes the $Q \times Q$ identity matrix, ε_g are i.i.d. component-specific errors with

Gaussian distribution $N(0,\Psi_g)$, where $\Psi_g = \text{diag}(\psi_{1g},\ldots,\psi_{pg})$, that are assumed to be independent of $\mathbf{F}_g$. Furthermore, β_{0g} and β_{1g} are the component-specific intercept and the $(1\times p)$ component-specific vector of the regression coefficients, respectively; while e_g is a component-specific disturbances variable with Gaussian distribution $N(0,\sigma_g^2)$. Moreover, by assuming that Y is conditionally independent of $\mathbf{F}$ given $\mathbf{X}=\mathbf{x}$ in the generic g-th mixture component, we get that the joint density of $(\mathbf{X},Y)$ is given by:

$$p(\mathbf{x},y,\theta) = \sum_{g=1}^{G} \pi_g N(y|\mathbf{x};m(\mathbf{x};\beta_g),\sigma_g^2) N(\mathbf{x};\mu_g,\Lambda_g\Lambda_g'+\Psi_g) \tag{2}$$

where $m(\mathbf{x};\beta_g) = \beta_{0g}+\beta_{1g}'\mathbf{X}$ and $\theta = \{\pi_g,\beta_g,\sigma_g^2,\Lambda_g,\Psi_g; g=1,\ldots,G\}$. A collection of sixteen parsimonious CWFA models can be obtained by constraining or not $\sigma_g^2=\sigma^2$, $\Lambda_g=\Lambda$, $\Psi_g=\Psi$, and$\Psi_g=\psi_g\mathbf{I}_p$.

3 The multivariate cluster-weighted disjoint factor analyzers model

As mentioned previously, here we introduce the MCWDFA model that extends CWFA framework by considering more than one response variable and by identifying factors that relate to disjoint clusters of explanatory variables which similarly predict the responses. Let $\mathbf{X}$ be the p-dimensional vector of explanatory variables and $\mathbf{Y}$ be the M-dimensional vector of the response variables. For each component g ($g=1,\ldots,G$), the MCWDFA model is composed of two parts. The first extends the regression model in (1) with a multivariate regression model formalizing the relations between the M responses and the p explanatory variables, as follows:

$$\mathbf{Y} = \mathbf{B}_{0g} + \mathbf{B}_{1g}'\mathbf{X} + \mathbf{e}_g \tag{3}$$

where $\mathbf{B}_{0g}$ and $\mathbf{B}_{1g}$ are the $(M\times 1)$ component-specific vector of intercepts and the $(p\times M)$ component-specific matrix of the regression coefficients, respectively; $\mathbf{e}_g$ is the $(M\times 1)$ component-specific vector of disturbances variables with Gaussian distribution $N(0,\Sigma_{\mathbf{e}_g})$. On the other hand, the second part of the model assumes that the factor loading structure of the CWFA model holds except for the factor loading matrix Λ_g. In fact, to introduce explanatory variable clustering forming disjoint clusters which similarly predict the responses, Λ_g is replaced by the product of the specific matrices $\mathbf{V}_g$ and $\mathbf{W}_g$, where $\mathbf{V}_g = [v_{jqg}]$ is a $(p\times Q)$ component-specific binary row stochastic matrix

representing the membership matrix of the explanatory variables into Q clusters corresponding to Q factors, i.e. $v_{jqg} = 1$ if and only if, for observations in the g-th component, the j-th explanatory variable belongs to cluster q, 0 otherwise ($j = 1, \ldots, p$); while, $\mathbf{W}_g = \text{diag}(w_{1g}, \ldots, w_{pg})$ is a $(p \times p)$ component-specific diagonal matrix of weights for the explanatory variables. Constraint $\mathbf{V}_g' \mathbf{W}_g \mathbf{W}_g \mathbf{V}_g = \text{diag}(w_{\cdot 1g}^2, \ldots, w_{\cdot Qg}^2)$, with $w_{\cdot qg}^2 = \sum_{j=1}^{p} w_{jqg}^2 > 0$ has to be satisfied, where the third index q added to w_{jg} indicates the factor associated with the j-th variable. Thus, the factor structure in (1) can be constrained in order to include the explanatory variables clustering as follows:

$$\mathbf{X} = \mu_g + \mathbf{W}_g \mathbf{V}_g \mathbf{F}_g + \boldsymbol{\varepsilon}_g. \tag{4}$$

It is interesting observe that, recalling similar factor assumptions of the CWFA model, the component-specific covariance matrix of $\mathbf{X}$, after the proper permutation of explanatory variables, has a block diagonal form, where each block is the component-specific covariance matrix of the subset of the explanatory variables related to a specific factor. Maximum likelihood parameter estimates are derived using an alternating expectation-conditional maximization (AECM) algorithm. Application of the proposed approach to both simulated and real datasets is presented.

References

BASILEVSKY, A. 1981. Factor analysis regression. *Canadian Journal of Statistics.*, **9**(1), 109–117.

GERSHENFELD, N. 1997. Nonlinear inference and cluster-weighted modeling. *Annals of the New York Academy of Sciences.*, **808**(1), 18–24.

INGRASSIA, S., MINOTTI, S.C., & VITTADINI, G. 2012. Local Statistical Modeling Via the Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification.*, **29**(3), 363–401.

MARTELLA, F., ALFO, M., & VICHI, M. 2008. Biclustering of gene expression data by an extension of mixtures of factor analyzers. *The international Journal of Biostatistics.*, **4**(1), 3.

SUBEDI, S., PUNZO, A, INGRASSIA, S., & MCNICHOLAS, P.D. 2013. Clustering and Classification Via Cluster-Weighted Factor Analyzers. *Advances in Data Analysis and Classification.*, **7**(1), 5–40.

VICHI, M. 2017. Disjoint factor analysis with cross-loadings. *Advances in Data Analysis and Classification.*, **11**(3), 563–591.

WEDEL, M. 2002. Concomitant Variables in Finite Mixture Models. *Statistica Neerlandica.*, **56**(3), 362–375.