FINITE MIXTURE MODELS FOR MIXED DATA: EM ALGORITHMS AND PARAFAC REPRESENTATIONS*

Marco Alfó 1, Paolo Giordani1

1 Dipartimento di Scienze Statistiche, Sapienza Università di Roma, (e-mail: marco.alfo@uniroma1.it, paolo.giordani@uniroma1.it)

1 Introduction

We propose a flexible regression model for multivariate mixed responses, where different numbers of locations are used for each margin, and joined by a full association structure, which can be simplified by using a Parafac-based representation. This structure of dependence is more general, and properly nests the independence model, but it does not come at no cost. In fact, by adopting this representation, the number of parameters can be shown to increase exponentially with the number of the analysed outcomes. Therefore, we propose a parsimonious representation of this multi-way array through the Parafac model, Harshman, 1970 and Kapteyn et al., 1986. This helps us define a flexible model that can account for profile-specific heterogeneity, general form of dependence between profiles and a parsimonious representations of the latter.

2 Modelling multivariate responses

Multivariate discrete responses have raised great interest in the last few years. In particular, latent variable models for mixed responses have received much attention; see Alfó & Rocchetti, 2013 for a discussion.

Let us assume we have observed the realizations $y_{ij}$ for $j = 1,\ldots,p$ outcomes on $i = 1,\ldots,n$ units, with a set of covariates $x_{ij}$, where $x_{ij1} \equiv 1$. When a reference multivariate distribution cannot be used to describe the dependence, latent variable models represent a typical choice. In this framework, we rely on conditional independence given (individual-specific) latent characteristics, and marginal dependence arises since the measurements from the same individual share some common latent features. Let $b_{ij}, i = 1,\ldots,n, j = 1,\ldots,p$, denote the set of individual- and outcome-specific random parameters. We assume $Y_{ij} | b_{ij}, x_{ij} \sim f(\theta_{ij})$, where $f(\cdot)$ is a member of the exponential family of

*The FIRB research project “Mixture and latent variable models for causal inference and analysis of socio-economic data” is gratefully acknowledged for financial support
distributions with canonical parameter:

\[ \theta_{ij} = x'_{ij} \beta_j + z'_{ij} b_{ij}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p. \]  

(1)

where \( z_{ij} \subseteq x_{ij} \). Here, \( \beta_j \) is an outcome-specific vector of (constant) regression parameters, while the random parameter vector \( b_i = (b_{i1}, \ldots, b_{ip}) \) has a multivariate density \( g(\cdot) \), with \( E(b_i) = 0 \). Under conditional independence:

\[
L(\cdot) = \prod_{i=1}^{n} \left\{ \int_{\mathcal{U}} \left[ \prod_{j=1}^{p} f(y_{ij} \mid x_{ij}, b_{ij}) \right] g(u_i) du \right\}.
\]

(2)

The marginal likelihood can not, in the general case, be written in closed form; to obtain maximum likelihood estimates, we may use numerical integration, see Rabe-Hesketh et al., 2005, or simulation-based techniques, see Chib & Winkelmann, 2001. An alternative is to leave the random parameter distribution unspecified and estimate it through a discrete distribution on \( K \leq n \) support points. Let \( \pi_g \) denote the mass associated to location \( b_g = (b_{g1}, \ldots, b_{gp}) \), \( g = 1, \ldots, K \). The likelihood function is:

\[
L(\cdot) = \prod_{i=1}^{n} \left\{ \sum_{g=1}^{K} \pi_g \left[ \prod_{j=1}^{p} f(y_{ij} \mid x_{ij}, b_{gj}) \right] \right\},
\]

(3)

Let us define the component indicators \( z_i, i = 1, \ldots, n \) and \( g = 1, \ldots, K \), with \( z_{ig} = 1 \) if the \( i \)-th unit comes from the \( g \)-th component, \( \pi_g = \text{Pr}(z_{ig} = 1) \). The EM algorithm, McLachlan & Krishnan, 1997 arises quite naturally in this context.

3 The proposal

While the finite mixture approach is more computationally efficient and robust than parametric approaches, it relies on unidimensionality assumption. We may notice by looking at eq. (3) that the model under independence does not occur as a special case of the dependence model. The log-likelihood for each outcome may refer to a different density and, therefore, be associated to a different weight in building up the global log-likelihood. We propose to adopt a dependence model that properly nests the independence one and a different number of components may be used for each outcome. Let us consider the following marginal representation \( P_j = \left\{ b_{gj}^{(j)}, \pi_{gj}^{(j)} \right\} \), where \( \pi_{gj}^{(j)} = \text{Pr}(b_{ij} = b_{gj}^{(j)}) \),
Marginals $\pi(j)$ control for heterogeneity in the univariate profiles, joint probabilities $\pi_{g_1,\ldots,g_p}$ describe the association between latent parameters across profiles. The following constraints hold:

$$\sum_{g_j=1}^{K_j} \pi_{g_j} = \sum_{g_1,\ldots,g_p} \pi_{g_1,\ldots,g_p} = 1.$$ 

Thus, $\prod_j K_j - 1$ mass parameters have to be estimated and this approach is as complex as those based on Gaussian quadrature. The number of locations is outcome-specific and this may help avoid to consider the same high number of locations for each profile. Further, the locations and prior probabilities are not fixed a priori. The log-likelihood function is:

$$\ell(\cdot) = \sum_{i=1}^{n} \log \left\{ \sum_{g_1,\ldots,g_p} \pi_{g_1,\ldots,g_p} \prod_j \left[ f\left(y_{ij} \mid x_{ij}, b_{g_j}\right)\right]\right\}$$

The choice for $K_j, j = 1,\ldots,p$ can be based on penalized likelihood criteria.

4 The use of the Parafac

The mass parameters $\pi_{g_1,\ldots,g_p}$ define a $p$-way array $\Pi$ with dimension $(K_1 \times \ldots \times K_g \times \ldots \times K_p)$. We may consider $p$-way arrays as generalizations of standard matrices where we have $p = 2$ way (rows and columns) and $p = 2$ modes (units and variables); when $p > 2$, we talk about $p$-way $p$-mode arrays. To summarize matrices, the widely-used Principal Component Analysis (PCA) can be applied, while the most famous multi-way extension of PCA is probably the Parafac model firstly proposed by Harshman, 1970 for the $p = 3$ case. The general $p$-way Parafac model, see Kapteyn et al., 1986, is

$$X = \sum_{s=1}^{q} a_1^{(1)} \circ \ldots \circ a_j^{(j)} \circ \ldots \circ a_p^{(p)} + E,$$ 

where $X$ denotes a $p$-way $p$-mode array of order $(K_1 \times \ldots \times K_p \times \ldots \times K_p)$, $A^{(j)}$ is the component matrix for the $j$-th way, $j = 1,\ldots,p$, with generic column $a_s^{(j)}$, $s = 1,\ldots,q$, $q$ denotes the number of extracted components, and $E$ is the error array. Finally, $\circ$ denotes the outer product. The idea is to reduce model complexity by using a Parafac model for $\Pi$. This amounts to moving
from $\prod_j K_j - 1$ to $(\sum_j K_j)q - (p - 1)q$. It is useful to remind that, under mild conditions, the Parafac solution is unique up to rescaling and joint permutation of the columns of the component matrices. To select the optimal number of components $q$, a penalized likelihood criterion can be used; the number of free parameters of the Parafac is lower whenever the order of one mode is remarkably higher than the others, for instance when $K_1 > K_2 \times \ldots \times K_p$.

Parameter estimation for the Parafac is usually carried out by minimizing the squared sum of the error terms. This can be done by Alternating Least Squares (ALS) algorithms, which update every component matrix separately, the remaining matrices held fixed. This means we have to iteratively solve $p$ different regression problems until convergence; in the present framework, the standard ALS algorithm cannot be used because the (log) mass parameters $\log(\pi_{g_1, \ldots, g_p})$ enters the expected value of the complete data loglikelihood function with weights given by posterior probabilities of component membership. Therefore, a Weighted Least-Squares (WLS) algorithm needs to be defined; such an algorithm can be derived by looking at the proposals developed, for the three-way case, by Kiers, 1997 and Bro et al., 2002. An iterative procedure can be considered where every step consists of applying the standard ALS algorithm for Parafac to a suitable transformation of the original data array. The approach is based on the minimization of a loss function that majorizes the WLS loss function, which is defined as minus the loglikelihood function (with unit constraint activated). This step can be embedded in the usual M step of an EM algorithm where, given the posterior probabilities of component membership, we derive updated estimates for model parameters and prior probabilities.

References


