

Finite mixtures for multivariate mixed data: a Parafac-based approach

Mixture finite per dati multivariati misti: un approccio basato sul Parafac

Paolo Giordani

Abstract We propose a flexible regression model for multivariate mixed responses. Discrete, outcome-specific, latent effects are used to account for potential dependence between outcomes. For this purpose, we define a multidimensional approach, where possibly different numbers of locations are used for each margin, and joined by a full association structure. This structure may be further simplified by using a *Parafac*-based approach.

Abstract *In questo lavoro, si descrive un modello di regressione flessibile per risposte multivariate di tipo misto. Sono utilizzati effetti latenti discreti, outcome-specifici, al fine di considerare potenziali strutture di dipendenza tra gli outcome. Per questo motivo, si definisce un approccio multidimensionale, nel quale un numero differente di locazioni può essere utilizzato per ciascun margine ed unito mediante una struttura di associazione completa. Tale struttura può successivamente essere semplificata utilizzando un modello di tipo Parafac.*

Key words: Finite mixtures of models, Latent variables, Parafac

1 Introduction

Multivariate discrete responses have raised great interest in the last few years. In particular, random effect models for mixed responses have received much attention, see, among others, [6], [3], [11], [2]. Among others, finite mixture models provide a simple way to implement semiparametric version of more computationally intensive methods based on Gaussian quadrature techniques. However, see eg [2], the finite mixture approach is based on the assumption of unidimensionality of the adopted latent structure, with the same components describing heterogeneity within (and de-

Paolo Giordani

Sapienza Università di Roma, P.le Aldo Moro, 5, 00185 Roma, Italy e-mail: paolo.giordani@uniroma1.it

pendence between) margins. therefore, no dependence is possible if heterogeneity is not present. So, it could be questioned whether this latent structure represents univariate heterogeneity, multivariate dependence or both. According to the approach in [2], we propose a general representation for the multivariate distribution of the random effects, where outcome specific heterogeneity is separated from the dependence between outcomes. The paper is structured as follows: in Section 2 we review the standard finite mixture approach to modeling multivariate responses. In Section 3 we show how this approach can be extended to more general structures of dependence when p responses of mixed type are observed. Section 4 describes how to apply the Parafac model to reduce the number of free parameters in the adopted model. Concluding remarks are given in Section 5.

2 Standard finite mixture approach

Let us assume we have observed responses y_{ij} , on $i = 1, \dots, n$ units and $j = 1, \dots, p$ outcomes, together with a set of (possibly outcome-specific) covariates $\mathbf{x}_{ij}$, where we assume that $x_{ij1} \equiv 1$. To model association between outcomes, we assume they share some common, unobservable, features. Let u_{ij} , $i = 1, \dots, n$, $j = 1, \dots, p$, denote the set of individual- and outcome-specific random effects; they account for heterogeneity across units in univariate profiles and dependence between unit-specific multivariate sequences. As it is standard in latent effect models, the observed responses y_{ij} are assumed to be conditionally independent (given the covariates and the random effects) random variables. Usually, the conditional density of the responses is a member of the exponential family

$$Y_{ij} \mid u_{ij}, \mathbf{x}_{ij} \sim \text{EF}(\theta_{ij})$$

and canonical parameters θ_{ij} are defined by the regression model:

$$\theta_{ij} = \mathbf{x}_{ij}^\top \beta_j + u_{ij}, \quad i = 1, \dots, n, \quad j = 1, \dots, p. \quad (1)$$

As it is clear, responses of different types from the exponential family can be modeled using this approach. Here, β_j is an outcome-specific vector of fixed regression parameters, while the random effect vector $\mathbf{u}_i = (u_{i1}, \dots, u_{ip})$ has a multivariate density $g(\cdot)$, with $E(\mathbf{u}_i) = \mathbf{0}$. By assuming conditional independence, the likelihood function is defined as follows:

$$L(\cdot) = \prod_{i=1}^n \left\{ \int_{\mathcal{U}} \left[\prod_{j=1}^p f(y_{ij} \mid \mathbf{x}_{ij}, u_{ij}) \right] g(\mathbf{u}_i) d\mathbf{u}_i \right\}. \quad (2)$$

For Gaussian random effects, the marginal likelihood can not, at least in the general case, be written in closed form; to obtain ML estimates, we may use numerical integration techniques such as standard or adaptive Gaussian Quadrature, see [11]

for a discussion. We may also rely on Monte Carlo or simulation-based techniques, see [6], [10]. However, marginal maximization using Gaussian quadrature or Monte Carlo approximations can be computationally very intensive; a potential alternative is based on leaving the random effect distribution unspecified and estimating it through a discrete distribution on $K \leq n$ support points. Let this distribution put masses π_g on locations $\mathbf{u}_g = (u_{g1}, \dots, u_{gp})$, $g = 1, \dots, K$, where u_{gj} represents the g -th location in the j -th profile, $g = 1, \dots, K$, $j = 1, \dots, p$. The resulting likelihood function is:

$$L(\cdot) = \prod_{i=1}^n \left\{ \sum_{g=1}^K \pi_g \left[\prod_{j=1}^p f(y_{ij} \mid \mathbf{x}_{ij}, u_{gj}) \right] \right\}, \quad (3)$$

where $\pi_g = \Pr(\mathbf{u}_g)$. Let us define the component indicators $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$, $i = 1, \dots, n$, with $z_{ig} = 1$ if the i -th unit comes from the g -th component. We may think at these component indicators as having a multinomial distribution with probabilities given by π_g , $g = 1, \dots, K$. The complete data likelihood is:

$$L_c(\cdot) = \prod_{i=1}^n \prod_{g=1}^K [f_{ig} \pi_g]^{z_{ig}}, \quad (4)$$

where $f_{ig} = \prod_{j=1}^p f_{ijg} = \prod_{j=1}^p f(y_{ij} \mid \mathbf{x}_{ij}, u_{gj})$, $i = 1, \dots, n$, $g = 1, \dots, K$. Since we do not observe the component indicators, the EM algorithm arises quite naturally.

3 A flexible approach

While the finite mixture approach is computationally efficient when compared to parametric ML approaches, it is based on a unidimensional latent variable, describing the unknown component the unit has been drawn from. This may lead to problems when the task is testing for dependence between the outcome-specific random effects, or when several outcomes are considered. As to the first point, it can be easily noticed by looking at eq. (3) that the model under independence does not occur as a special case of the dependence model. In fact, the marginal distributions under independence are estimated via univariate random effect models that are based on a number of components that may differ when compared across outcomes and with the dependence model. As to the second point, we should bear in mind that 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. Therefore, the *estimated* number of components may be driven by a subset of the analyzed outcomes only. For all these reasons, we propose to adopt a different parameterization for the prior probabilities; by using this approach, the dependence model properly nests the independence one and a different number of components may be used for each outcome. Let us adopt the following marginal representation for the outcome specific random effects:

$$P_j = \left\{ u_{g_j}^{(j)}, \pi_{g_j, j} \right\}_{g_j=1, \dots, K_j},$$

where $\pi_{g_j} = \Pr(u_{ij} = u_{g_j}^{(j)})$, $g_j = 1, \dots, K_j$, $j = 1, \dots, p$, and

$$\pi_{g_j} = \sum_{g_1, \dots, g_{j-1}, g_{j+1}, \dots, g_p} \pi_{g_1, \dots, g_p}.$$

Note that we use $\sum_{g_1, \dots, g_p} \pi_{g_1, \dots, g_p}$ as a shorthand for $\sum_{g_1=1}^{K_1} \dots \sum_{g_p=1}^{K_p} \pi_{g_1, \dots, g_p}$. The

mass $\pi_{g_1, \dots, g_p}$ is associated to each p -uple of random effects, say $(u_{g_1}^{(1)}, \dots, u_{g_p}^{(p)})$, $g_j = 1, \dots, K_j$, $j = 1, \dots, p$, and the number of components K_j may differ across outcomes. While marginals control for heterogeneity in the univariate profiles, joint probabilities describe the association between the latent effects in the p profiles. This approach can be considered as a standard finite mixture approach with $K = \prod_j K_j$ components, where each of the K_j locations in the j -th profile appears in a p -ple with each of the $\prod_{l \neq j} K_l$ locations in the remaining $p - 1$ profiles. Obviously, when $p = 1$ the proposed model reduces to a standard univariate finite mixture model. The following constraints hold:

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

Thus, $\prod_j K_j - 1$ mass parameters have to be estimated. This means that, from a computational point of view, this approach is approximately as complex as those based on Gaussian quadrature; however, we may point out, at least, two differences. First, the number of locations is outcome-specific and this may help save computational times, as we may avoid to consider a high number of locations for profiles with low heterogeneity. Second, the locations and prior probabilities are not constrained to represent the discretization of a standard Gaussian distribution. Therefore, for a given heterogeneity, the number of locations may be lower than the number of Gaussian quadrature abscissas (usually at least 7 for each profile). The log-likelihood function is:

$$\begin{aligned} \ell(\cdot) &= \sum_{i=1}^n \log \left\{ \sum_{g_1, \dots, g_p} \pi_{g_1, \dots, g_p} \prod_j \left[f(y_{ij} \mid \mathbf{x}_{ij}, u_{g_j}^{(j)}) \right] \right\} \\ &= \sum_{i=1}^n \log \left\{ \sum_{g_1, \dots, g_p} \pi_{g_1, \dots, g_p} f(\mathbf{y}_i \mid \mathbf{x}_{i1}, \dots, \mathbf{x}_{is}, u_{g_1}^{(1)}, \dots, u_{g_p}^{(p)}) \right\} = \sum_{g_1, \dots, g_p} \pi_{g_1, \dots, g_p} f_{i, g_1, \dots, g_p}. \end{aligned}$$

Once the choice for K_j , $j = 1, \dots, p$ has been made, for example through penalized likelihood criteria, eg BIC, [12], CAIC, [5], or ICL, [4], the EM algorithm is quite simple to be implemented, by adding factor variables with K_j levels in each model. We choose to use the AIC [1] to compare models with different number of

components, since it produces a more refined estimate of the mixing distribution, as shown by [9].

4 Flexibility through the Parafac model

To reduce the number of free parameters in the $\pi_{g_1, \dots, g_p}$'s, the so-called Parafac model [7] can be used. The Parafac represents the most natural extension of standard Principal Component Analysis to arrays with an arbitrary number of ways ('unit by variable' matrices can be seen as arrays with two ways and two modes, i.e. units and variables). Let the probabilities $\pi_{g_1, \dots, g_p}$'s be the elements of a p -way array denoted by $\underline{\Pi}$. The idea is to summarize such probabilities by a limited number of latent components that approximate the original probabilities at best.

Let $\mathbf{A}^{(j)}$ be the component matrix for the j -th way, $j = 1, \dots, p$, with generic column $\mathbf{a}_s^{(j)}$, $s = 1, \dots, q$. The Parafac model for a p -way p -mode array can be formalized as

$$\underline{\mathbf{X}} = \sum_{s=1}^q \mathbf{a}_s^{(1)} \circ \dots \circ \mathbf{a}_s^{(j)} \circ \dots \circ \mathbf{a}_s^{(p)} + \underline{\mathbf{E}}, \quad (5)$$

where $\underline{\mathbf{X}}$ is the generic p -way p -mode data array of order $(K_1 \times \dots \times K_g \times \dots \times K_p)$ with generic element $x_{g_1, \dots, g_j, \dots, g_p}$, $\underline{\mathbf{E}}$ is the error array and $\circ$ denotes the outer product. The model in (5) is a particular case of the more general one proposed in [8]. The optimal component matrices are estimated in the least-squares sense by minimizing the squared sum of the error terms; for this purpose, an Alternating Least Squares (ALS) algorithm can be adopted. In ALS algorithms, each parameter matrix is updated while keeping the remaining parameter matrices fixed; in the Parafac case, all the updates can be obtained by solving standard regression problems. Starting from the findings for the three-way case [13], the number of free parameters of (5) is $(K_1 + \dots + K_p)q - (p-1)q$. Such a number depends on the sum of the number of elements of the component matrices for all the modes. The last term corrects for the scaling freedom, where scaling the component matrices of $p-1$ modes fixes the scaling of the last one. In fact, it is worth recalling that, under mild assumptions, the Parafac solution is unique up to rescaling and joint permutation of the columns of the component matrices. Also note that the number of free parameters is lower when the number of entities of one mode is higher than those of the other modes. For instance, in the three-way case, if $K_1 > K_2 K_3$, then the data array can be reduced to an array of order $(K_2 K_3 \times K_2 \times K_3)$ and the number of free parameters can be computed replacing K_1 by $K_2 K_3$.

In the present context, we have p random effects and the Parafac model reported in (5) is used to approximate the p -way probability array $\underline{\Pi}$. It is easy to see that the use of Parafac to summarize $\underline{\Pi}$ reduces the number of free parameters from $\prod_j K_j - 1$ to $(\sum_j K_j)q - (p-1)q$. The AIC can be considered for choosing the best value of q , the number of *prototypes* used to approximate $\underline{\Pi}$.

5 Final remarks

A flexible regression model for multivariate mixed responses has been proposed. The goal is to introduce a general representation for the multivariate distribution of the random effects. To tune the complexity of such a general representation, we propose to use a Parafac model. Some results will be presented at the meeting.

References

1. Akaike, H.: A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19**, 716–723 (1974)
2. Alfò, M., Rocchetti, I.: A flexible approach to finite mixture regression models for multivariate mixed responses. *Statistics and Probability Letters* **83**, 1754–1758 (2013)
3. Alfò, M., Trovato, G.: Semiparametric mixture models for multivariate count data, with application. *Econometrics Journal* **7**, 1–29 (2004)
4. Biernacki, C., Celeux, G., Govaert, G.: Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 719–725 (2000)
5. Bozdogan, H.: Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika* **52**, 345–370 (1987)
6. Chib, S., Winkelmann, R.: Markov chain monte carlo analysis of correlated count data. *Journal of Business and Economic Statistics* **19**, 428–435 (2001)
7. Harshman, R.A.: Foundations of the PARAFAC procedure: Models and conditions for an 'exploratory' multimode factor analysis. *UCLA Working Papers in Phonetics* **16**, 1–84 (1970)
8. Kapteyn, A., Neudecker, H., Wansbeek, T.: An approach to n -mode components analysis. *Psychometrika* **51**, 269–275 (1986)
9. Karlis, D., Meligkotsidou, L.: Finite mixtures of multivariate poisson distributions with application. *Journal of Statistical Planning and Inference* **137**, 1942–1960 (2007)
10. Munkin, M., Trivedi, P.: Simulated maximum likelihood estimation of multivariate mixed-poisson regression models, with application. *Econometrics Journal* **2**, 29–48 (1999)
11. Rabe-Heskett, S., Skrondal, A., Pickles, A.: Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects. *Journal of Econometrics* **128**, 301–323 (2005)
12. Schwarz, G.E.: Estimating the dimension of a model. *Annals of Statistics* **6** 461–464 (1978)
13. Wansbeek, T., Verhees, J.: Models for multidimensional matrices in econometrics and psychometrics. In: Coppi, R., Bolasco, S. (eds.) *Multiway data analysis*, pp. 543–552. North Holland, Amsterdam (1989)