Bayesian Flexible Modelling of Mixed Logit Models

Luisa Scaccia[1] and Edoardo Marcucci[2]

[1] Dip. di Istituzioni Economiche e Finanziarie, Università di Macerata
via Crescimbeni 20, 62100 Macerata, Italy, *scaccia@unimc.it*
[2] Dip. di Istituzioni Pubbliche, Economia e Società, Università di Roma Tre
Via G. Chiabrera 199, 00145 Roma, Italy, *edoardo.marcucci@uniroma3.it*

Abstract. The widespread use of the Mixed Multinomial Logit model, in the context of discrete choice data, has made the issue of choosing a mixing distribution very important. The choice of a specific distribution may seriously bias results if that distribution is not suitable for the data. We propose a flexible hierarchical Bayesian approach in which the mixing distribution is approximated through a mixture of normal distributions. Numerical results on a real data set are provided to demonstrate the usefulness of the proposed method.

Keywords: Hierarchical Bayes, mixed logit, mixture of distributions, random taste heterogeneity, semiparametric estimation

1 Introduction

The multinomial logit (MNL) model has provided for a long time the foundation for the analysis of discrete choice modelling, due to its advantages in terms of closed-form solution and simplicity of interpretation and use (McFadden (1974)). However, some restrictive assumptions underlying the model have motivated researchers to consider alternative specifications, the most popular of which is probably the mixed logit (MMNL) model (McFadden and Train (2000), Train (1998)). In its simplest specification, the utility of each individual is a function of the alternative attributes, with attribute coefficients that are random and reflect individual preferences.

In MMNL models, however, a crucial issue is that of specifying an appropriate mixing distribution of the random coefficients that may be interpreted as representing random taste heterogeneity. Most popular specifications have been the normal, triangular, uniform and lognormal distributions. However, in practical applications, any of them has shown its deficiencies (Hess et al. (2005)). An unappropriate choice of the mixing distribution can lead to problems in interpretation and potentially misguided policy-decisions (Cirillo and Axhausen (2006), Fosgerau (2006)).

To deal with this issue, Fosgerau and Hess (2009) proposed two approaches: the first one improves on the flexibility of a base distribution by adding in a series approximation using Legendre polynomials; the second one

makes use of a semi-parametric mixing distribution consisting of a discrete mixture of normal distributions (MOD). Both approaches can approximate any continuous distribution, allowing also for multiple modes, a significant advantage compared to typically used distributions, restricted to a single mode. Allowing for multiple modes means that the population may be composed of distinct groups with different behaviour. In a Monte Carlo study, Fosgerau and Hess (2009) show that the two approaches do about equally well in outperforming commonly used distributions, over a range of situations. The MOD approach has a particular ability in approximating point masses. A heightened mass at zero is useful in representing taste heterogeneity for attributes that some individuals are indifferent to, as discussed by Cirillo and Axhausen (2006) with regard to valuation of travel time savings.

In this paper, we consider the MOD approach and we illustrate how to estimate this model in a Bayesian framework. Moreover, we extend the approach to the case in which multiple random coefficients, potentially correlated, are present in the model. As Fosgerau and Hess (2009), we will fix the number of components in the mixture. The extension to mixtures allowing the number of components to vary is a topic for further research.

We rely on Bayesian procedures since these avoid two of the most prominent difficulties associated with classical procedures. Firstly, the Bayesian procedures do not require maximization of any function, thus avoiding the numerical difficulties that often arise in maximizing the simulated likelihood function of some MMNL models. Secondly, desirable estimation properties, such as consistency and efficiency, can be attained under more relaxed conditions with Bayesian procedures (Train (2001)). Moreover, Bayesian procedures usually avoid the need to simulate choice probabilities, which is quite cumbersome with MMNL models. Finally, individual-level parameters can be easily obtained. The Bayesian perspective has been adopted in the context of discrete choice models by, for example, Train (2001) for mixed logits with normal, lognormal, uniform and triangular distributed coefficients. Allenby et al. (1998) used a mixture of normal distribution for random parameters in mixed logits, in the context of marketing research. We extend their approach, including fixed parameters in the model and allowing for a more flexible hierarchical structure. Our approach also relates to the one proposed by Ho and Hu (2008) for linear mixed models with random effects.

The paper is organized as follows: the model and prior assumptions are illustrated in Section 2; Section 3 deals with computational implementation; Section 4 discusses an application to the analysis of stated preference data on public transport demand. Conclusions are given in Section 5.

2 The Mixed logit model

In this Section, we illustrate the MOD approach in a hierarchical Bayesian fashion. We, then, specify priors for the parameters in the model.

2.1 The MOD approach

Person n faces a choice among J alternatives in each of T time periods. According to the specification of a MMNL model that we will use, the person's utility from alternative i in period t is:

$$U_{nit} = \alpha' w_{nit} + \beta_n' x_{nit} + \epsilon_{nit} \quad (n = 1, \ldots, N;\ i = 1, \ldots, J;\ t = 1, \ldots, T)$$

where $\epsilon_{nit} \sim$ i.i.d. extreme value, w_{nit} is a vector of R attributes (characterizing the alternative and/or the subject) whose coefficients α are fixed and x_{nit} is a vector of K attributes whose coefficients β_n are supposed to be random and to vary in the population, according to the density $g(\beta_n|\mu, \Sigma)$, for $n = 1, \ldots, N$, with μ and Σ being hyperparameters. Person n chooses alternative i in period t if $U_{nit} > U_{njt}$, $\forall j \neq i$. Let $y_n = (y_{n1}, \ldots, y_{nT})$ be the person's sequence of choices over the T time periods. The probability of observing this sequence, conditional on the person-specific parameters β_n and the common fixed parameters α, is the product of standard logit formulas:

$$L\left(y_n|\alpha, \beta_n\right) = \prod_{t=1}^{T} \frac{e^{\alpha' w_{ny_{nt}t} + \beta_n' x_{ny_{nt}t}}}{\sum_{j=1}^{J} e^{\alpha' w_{njt} + \beta_n' x_{njt}}}. \tag{1}$$

The MOD approach, adopting a semi-parametric perspective, assumes that $g(\beta_n|\mu, \Sigma)$ is a mixture of C multivariate normal distributions:

$$\beta_n|\mu, \Sigma \sim \sum_{c=1}^{C} s_c \phi(\cdot|\mu_c, \Sigma_c) \qquad (n = 1 \ldots, N), \tag{2}$$

where $\phi(\cdot|\mu_c, \Sigma_c)$ is a multivariate normal density with mean vector $\mu_c = (\mu_{c1}, \ldots, \mu_{cK})'$ and K by K covariance matrix Σ_c, and s_c are weights satisfying $0 \leq s_c \leq 1$, for $c = 1, \ldots, C$, and $\sum_{c=1}^{C} s_c = 1$.

Notice that this model provides both the flexibility of the latent class model and the parsimony of the traditional MMNL model. Indeed, both models are special cases of the proposed model: the latent class model is obtained by letting the within-class variances go to zero, and the traditional MMNL model corresponds to using only one class or component.

2.2 Latent allocation variables

An alternative perspective, leading to the same mixture model in (2), involves the introduction of latent allocation variables $z = (z_1, \ldots, z_N)$ and the assumption that the vector β_n, relative to individual n, arose from an unknown component z_n of the mixture of multivariate normal distributions. The allocation variables are given probability mass function

$$p(z_n = c) = s_c \qquad \text{independently for } n = 1, \ldots, N, \tag{3}$$

and conditional on them, the random taste parameters β_n are independently drawn, for each subject n, from the density:

$$\beta_n | z, \mu, \Sigma \sim \phi(\cdot | \mu_{z_n}, \Sigma_{z_n}). \tag{4}$$

Integrating out z_n in (4), using the distribution in (3), leads back to (2).

2.3 Prior settings

We assume the number of components C to be fixed to a reasonable small number, as in Fosgerau and Hess (2009). In a forthcoming paper, we will consider C to be unknown and subject to inference, as well as the other parameters of the model. From past experience, we would not expect inference about the model proposed to be highly sensitive to prior specification. We use a weakly informative priors approach, according to which we use some information from the sample to set the values of the hyperparameters. In particular, we fit a standard mixed logit model to the data, with normal distribution of the taste parameters to get some idea from the estimated mean and standard errors of the random parameters (Ho and Hu (2008)).

In particular, we assume a priori:

a) $(s_1, \ldots, s_C) \sim \mathcal{D}(\delta, \ldots, \delta)$, where $\mathcal{D}$ denotes the Dirichlet distribution. We choose $\delta = 1$.
b) $z_n \sim p(z_n = c) = s_c$, independently for $n = 1, \ldots, N$.
c) $\Sigma_c \sim \mathcal{IW}(r, \Theta^{-1})$, independently for $c = 1, \ldots, C$, where $\mathcal{IW}$ denotes the Inverse Wishart distribution.
d) $\mu_c \sim \mathcal{N}_K(\xi, D)$, independently for $c = 1, \ldots, C$, where $\mathcal{N}_K$ denotes the K-dimensional multivariate normal distribution.
e) $\Theta \sim \mathcal{IW}(a, S^{-1})$.
f) $\alpha \sim \mathcal{N}_R(\nu, \Omega)$.

2.4 Complete hierarchical model

Let $y = (y_1', \ldots, y_N')'$, $\mu = (\mu_1', \ldots, \mu_C')'$ and Σ be the matrix obtained by stacking the covariance matrices Σ_c on top of each other. We exploit the natural conditional independence structure so that the joint distribution of all variables, conditional to the fixed values of the hyperparameters, is

$$\begin{aligned} p(y, s, z, \Sigma, \Theta, \mu, \alpha, \beta | C, \nu, \gamma, \xi, D, a, S, r, \delta) \\ = p(s|C, \delta) p(z|s, C) p(\Theta|a, S) p(\Sigma|r, \Theta, C) \\ \cdot p(\mu|\xi, D, C) p(\alpha|\nu, \Omega) p(\beta|z, \mu, \Sigma) p(y|\alpha, \beta), \end{aligned}$$

where $p(z|s, C) = \prod_{n=1}^{N} s_{z_n}$, $p(y|\alpha, \beta) = \prod_{n=1}^{N} L(y_n|\alpha, \beta_n)$, with $L(y_n|\alpha, \beta_n)$ defined in (1) and $p(\beta|z, \mu, \Sigma)$ given in (4). The prior distributions $p(s|C, \delta)$, $p(\Theta|a, S)$, $p(\Sigma|r, \Theta, C)$, $p(\mu|\xi, D, C)$, $p(\alpha|\nu, \Omega)$ are all given in Section 2.3.

3 Computational implementation

The complexity of the model presented requires Markov chain Monte Carlo (MCMC) methods to approximate the posterior distribution. Our sampler uses seven fixed-dimension moves. Gibbs samplers are used to update all model parameters, except α and β, which are updated by means of Metropolis algorithm. The notation '$\cdots$' will be used to denote 'all other variables'.

Updating s. Before considering the updating of s, we comment briefly on the issue of labeling the components. The whole model is, in fact, invariant to permutation of the labels $c = 1, \ldots, C$. For identifiability, we adopt a unique labeling in which the component weights are in increasing numerical order. As a consequence, the joint prior distribution of s is a Dirichlet density, restricted to the set $s_1 < s_2 < \ldots < s_C$. The weights are updated by drawing them from their full conditional distribution

$$(s_1, \ldots, s_C)|\cdots \sim \mathcal{D}(\delta + m_1, \ldots, \delta + m_C)$$

where $m_c = \#\{n : z_n = c\}$ is the number of subjects currently allocated to the c component of the mixture. To preserve the ordering constraints on s, the move is accepted provided the ordering is unchanged.

Updating z. The allocation variable z_n has conditional probability

$$p(z_n = c|s, C, \beta_n) = \frac{s_c \phi(\beta_n|\mu_c, \Sigma_c)}{\sum_{c=1}^{C} s_c \phi(\beta_n|\mu_c, \Sigma_c)}.$$

We can update the z_n independently, sampling from this distribution.

Updating μ. The μ_c can be updated independently, drawing them from

$$\mu_c|\cdots \sim \mathcal{N}_K\left(\frac{D^{-1}\xi + m_c\Sigma_c^{-1}\bar{\beta}_c}{D^{-1} + m_c\Sigma_c^{-1}}, \frac{1}{D^{-1} + m_c\Sigma_c^{-1}}\right)$$

where $\bar{\beta}_c = m_c^{-1}\sum_{n:z_n=c}\beta_n$.

Updating Θ. We update Θ sampling from its full conditional:

$$\Theta|\cdots \sim \mathcal{IW}\left(a + Cr, S^{-1} + \sum_{c=1}^{C}\Sigma_c^{-1}\right)$$

Updating Σ. We update Σ_c independently, sampling from

$$\Sigma_c|\cdots \sim \mathcal{IW}\left(m_c + r, \Theta^{-1} + \sum_{n:z_n=c}(\beta_n - \mu_c)(\beta_n - \mu_c)'\right)$$

Updating α. The Metropolis algorithm to update α proposes, at step $h+1$, a new value $\alpha^\star$ drawn from a symmetric proposal density $\mathcal{N}_R(\alpha^{(h)}, \tau_1\Omega)$, where τ_1 is a tuning parameter. This proposal is accepted with probability

$$\min\left\{1, \frac{\prod_{n=1}^{N} L\left(y_n|\alpha^\star, \beta_n^{(h)}\right)\phi(\alpha^\star|\nu, \Omega)}{\prod_{n=1}^{N} L\left(y_n|\alpha^{(h)}, \beta_n^{(h)}\right)\phi(\alpha^{(h)}|\nu, \Omega)}\right\}.$$

If the proposal is accepted, $\alpha^{(h+1)} = \alpha^{\star}$, otherwise $\alpha^{(h+1)} = \alpha^{(h)}$.

Updating β. We update β_n independently, by means of Metropolis algorithm. At the $h+1$ step of the algorithm, we use a $\mathcal{N}_K(\beta_n^{(h)}, \tau_2 \Sigma_{z_n})$ as symmetric proposal density, with τ_2 being a tuning parameter, and we accept the new value $\beta_n^{\star}$, drawn from it, with probability

$$\min\left\{1, \frac{L\left(y_n|\alpha^{(h)}, \beta_n^{\star}\right)\phi(\beta_n^{\star}|\mu_{z_n}, \Sigma_{z_n})}{L\left(y_n|\alpha^{(h)}, \beta_n^{(h)}\right)\phi(\beta_n^{(h)}|\mu_{z_n}, \Sigma_{z_n})}\right\}.$$

4 An application to public transport demand

The data set refers to a study carried out in Urbino (Italy) to analyse the attributes of the local public transport and to investigate possible interventions to improve the service (Marcucci and Scaccia (2005), Scaccia (2010)). Five attributes of the service were considered: cost of monthly ticket, headway, first and last run, real time information displays, bus shelters. Each attribute was further described by five levels. Questionnaires contained 15 choice exercises, 11 of which were random, 2 aimed at testing the quality of the answers, and 2 aimed at testing preference stability. Each choice exercise contained four hypothetical alternatives. A total number of 50 respondents took part in the study, providing a data set of 750 observations.

To specify the models, the Lagrange multiplier test (McFadden and Train (2000)) was used to decide which parameters are to be random. The null hypothesis of no mixing was rejected for the parameters of the attributes headway and daily operating time. The cost parameter was treated as non random to simplify the estimation of marginal willingness to pay for an improvement in a certain attribute (see Scaccia (2010)).

To estimate the proposed model, we performed 100,000 sweeps of the MCMC algorithm, allowing for a burn-in of 50,000 sweeps. Posterior means of relevant parameters are given in Table 1. The posterior estimates for the fixed parameters are very close to those obtained by Scaccia (2010). The signs of the cost and bus shelters parameters are as expected, while the information displays attribute seems to have a non significant influence on utility.

The marginal posterior densities for the random parameters are shown in Figure 1. The estimated posterior distribution of β_{headway} does not change when moving from the MMNL model with normal mixing density specification to the MOD models. In this case, a simple normal mixing density would have been appropriate to approximate taste heterogeneity with respect to the headway attribute. This shows how the MOD approach can also be seen as a diagnostic tool to get an idea of the shape of the true distribution and to help in the choice of an appropriate model. The estimated posterior distribution of $\beta_{\text{run time}}$ is, instead, different under the three models. Under both the MOD models, a mass point at zero can be noticed, revealing the presence of

	MMNL		MOD (2 components)		MOD (3 components)	
Parameter	Est.	Std. dev.	Est.	Std. dev.	Est.	Std. dev.
α_{cost}	-0.2671	0.0336	-0.2751	0.0346	-0.2764	0.0349
α_{displays}	-0.0452	0.1753	-0.0190	0.1678	-0.0242	0.1828
α_{shelters}	0.3258	0.1771	0.3167	0.1835	0.3320	0.1747
μ_{headway}	-0.1039	0.0247	-0.0673	0.0806	-0.0485	0.1525
			-0.1116	0.0254	-0.0750	0.0864
					-0.1127	0.0265
$\mu_{\text{run time}}$	0.5322	0.0537	0.0403	0.1454	0.0180	0.1926
			0.6561	0.0798	0.0863	0.1697
					0.6840	0.0867
s	1.0000	0.0000	0.1599	0.0891	0.0503	0.0398
			0.8401	0.0891	0.1623	0.0818
					0.7873	0.0957

Table 1. *Posterior mean estimates of relevant parameters.*

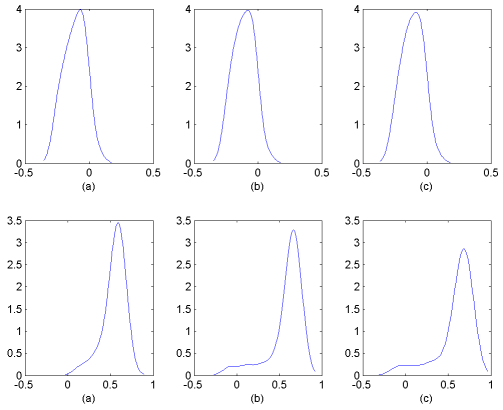

Fig. 1. Estimated marginal posterior densities for the random parameters β_{headway} (upper panel) and $\beta_{\text{run time}}$ (lower panel) under a) the MMNL model, b) the MOD model with 2 components, c) the MOD model with 3 components.

individuals that are indifferent to the availability of bus departures early in the morning or late at night. This interesting feature is not revealed by the MMNL model, which is not flexible enough to represent taste heterogeneity with respect to run time.

5 Conclusions and further development

We proposed modelling mixed logit models under a Bayesian hierarchical framework, making use of a mixture of normal distribution to approximate the density of the random parameters. This approach is conceptually simple and is flexible enough to approximate well a variety of distributions, allowing for multiple modes, as well as for saddle points in a distribution. Furthermore, it is possible to have point-mass at a specific value. Lastly, the mixture components can be used to classify the individuals into homogeneous groups, facilitating cluster analysis and classification. The approach proposed is not restricted to being based on the normal distribution but can use any continuous distribution. This could be an interesting avenue for further research.

References

ALLENBY, G.M., ARORA, N. and GINTER, J.L. (1998): On the heterogeneity of demand. *Journal of Marketing Research 35 (3), 384-389.*

CIRILLO, C. and AXHAUSEN, K.W. (2006): Evidence on the distribution of values of travel-time savings from a six-week travel diary. *Transportation Research Part A: Policy and Practice 40 (5), 444-457.*

FOSGERAU, M. (2006): Investigating the distribution of the value of travel time savings. *Transportation Research Part B: Methodological 40 (8), 688-707.*

FOSGERAU, M. and HESS, S. (2009): A comparison of methods for representing random taste heterogeneity in discrete choice models. *European Transport Trasporti Europei 42, 1-25.*

HESS, S., BIERLAIRE, M. and POLAK, J.W. (2005): Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice 39 (2-3), 221-236.*

HO, R.K.W. and HU, I. (2008): Flexible modelling of random effects in linear mixed models-A Bayesian approach. *Computational Statistics and Data Analysis 52 (3), 1347-1361.*

MARCUCCI, E. and SCACCIA, L. (2005): Alcune applicazioni dei modelli a scelta discreta al settore dei trasporti. In: E. Marcucci (Ed.): *I Modelli a Scelta Discreta nel Settore dei Trasporti. Teoria, Metodologia e Applicazioni*, Carocci editore, Roma.

McFADDEN, D. (1974): Conditional logit analysis of qualitative choice behaviour. In: P.C. Zarembka (Ed.): *Frontiers in Econometrics.* Academic Press, New York, 105–142.

McFADDEN, D. and TRAIN, K. (2000): Mixed MNL models for discrete response. *Journal of Applied Econometrics 15 (5), 447-470.*

SCACCIA, L. (2010): Random parameters logit models applied to public transport demand. *Forthcoming.*

TRAIN, K. (1998): Recreation demand models with taste differences over people. *Land Economics, 74 (2), 230-239.*

TRAIN, K. (2001): A Comparison of hierarchical bayes and maximum simulated likelihood for mixed logit. *Working Paper*, Department of Economics, University of California, Berkeley.