A Choice Model with Infinitely Many Latent Features

Dilan Görür DILAN@TUEBINGEN.MPG.DE
Frank Jäkel FRANK@TUEBINGEN.MPG.DE
Carl Edward Rasmussen CARL@TUEBINGEN.MPG.DE
Max-Planck-Institute for Biological Cybernetics, Department of Empirical Inference, Tübingen, Germany

Abstract

Elimination by aspects (EBA) is a probabilistic choice model describing how humans decide between several options. The options from which the choice is made are characterized by binary features and associated weights. For instance, when choosing which mobile phone to buy the features to consider may be: long lasting battery, color screen, etc. Existing methods for inferring the parameters of the model assume pre-specified features. However, the features that lead to the observed choices are not always known. Here, we present a non-parametric Bayesian model to infer the features of the options and the corresponding weights from choice data. We use the Indian buffet process (IBP) as a prior over the features. Inference using Markov chain Monte Carlo (MCMC) in conjugate IBP models has been previously described. The main contribution of this paper is an MCMC algorithm for the EBA model that can also be used in inference for other non-conjugate IBP models—this may broaden the use of IBP priors considerably.

1. Introduction

Psychologists have long been interested in the mechanisms underlying choice behavior (Luce, 1959). In virtually all psychological experiments subjects are asked to make a choice and the frequency of the responses is recorded. Often the choice is very simple like pressing one of two buttons. However, even in these simple choices one finds probabilistic responses. In what seem to be identical experimental conditions one can observe that different subjects respond differently. Responses vary even for the same subject that is repeatedly presented with the same choice scenario. In economics, too, choice is an omnipresent phenomenon: Consumers choose one brand instead of another, commuters choose to take the bus rather than the car and college students prefer one university over another. Considerable effort has been put into probabilistic modeling of data arising from such choice situations (McFadden, 2000; Train, 2003).

Appearing in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 2006.

In accordance with economic theory, it is often assumed that humans are rational and make choices by maximizing utility. The reason for the observed probabilistic variations in choice behavior is random variations in utility. These variations may arise because different decision makers make different judgments about the utility of an option, or because each decision maker varies randomly in her assessment of utility over time. Models that fit into this framework are called random utility models (RUMs). In contrast to RUMs many psychological models do not assume a rational decision maker—instead they attempt to explain the (probabilistic) mental processes that take place in the course of a decision.

The Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Luce, 1959), which can be seen both as a RUM and as a psychological process model, is one of the most influential models. However, it is well-known that the BTL cannot account for all of the choice patterns that can be observed in choice data. There is a large literature on how choice data can violate the assumptions built into the BTL model (Restle, 1961; Rumelhart & Greeno, 1971; Tversky, 1972; Train, 2003).

Within the RUM framework several other models have been suggested to account for these cases, e.g. the probit model with correlated noise, the nested logit and mixed logit models (Train, 2003). Recently, a mixed multinomial logit model with a non-parametric mixing distribution has been proposed by James and Lau (2004) using the Dirichlet process.

In psychology, the elimination by aspects (EBA) model—which includes the BTL model as a special case—is probably the most prominent model. In the EBA model it is assumed that options are represented by binary feature vectors, called aspects. If the choice was between several mobile phones the features could be whether they have a built-in MP3-player, the display is in color, the battery lasts long enough, etc. Each feature has a weight associated with it reflecting the importance of each aspect for the choice. The subject selects a feature at random (but more important features have a higher probability to be selected) and eliminates all options that do not have this feature. This process is repeated until only one option remains. If the features are known their weights can be estimated from choice data (Wickelmaier & Schmid, 2004). However, generally it is not known what the features are and one would like to infer them from observed choices. The usefulness of the EBA model for the analysis of choice data has been extremely limited by the fact that inference about the underlying features is very difficult (Tversky & Sattath, 1979).

We can treat the binary features as random and do inference on them. The number of features can be chosen by model selection. In a non-parametric Bayesian setting it is also possible to use infinitely many features. Recently, a prior over sparse binary matrices with infinitely many columns—the Indian buffet Process (IBP)—has been defined by Griffiths and Ghahramani (2005). The IBP has been used as a prior for the latent feature matrix in *additive clustering* by Navarro and Griffiths (2005). This suggests to use the IBP as a prior for the latent feature matrix in the EBA model. Analytical inference in this model is intractable and therefore we use Markov Chain Monte Carlo (MCMC) methods.

For the model considered by Navarro and Griffiths (2005) the prior distribution of the weights is chosen to be conjugate to the likelihood. Therefore, the conditional posteriors needed for inference using Gibbs sampling can be calculated analytically. However, our formulation of the EBA does not allow integration over the weights. This prevents us from calculating the posterior for the infinite feature matrix. We approximate the posterior by truncating the feature matrix and we use auxiliary variables for the weights.

We describe the non-parametric Bayesian formulation of the EBA model in the next section. The MCMC algorithm derived for inference in the non-conjugate IBP models is presented in section 3, followed by experimental results in section 4, and we conclude with a discussion in section 5.

2. Model Specification

The EBA model is defined for choice from a set of several options but for clarity of presentation we consider only the paired comparison case here which reduces the EBA model to Restle's choice model (Restle, 1961; Tversky, 1972). The inference techniques we describe below are also valid for the general EBA model.

In a paired comparison experiment there is a set of N options but the subjects are presented only with pairs of options at a time. The task of the subject is to indicate which of two options i and j she prefers. Options are described by K-dimensional binary feature vectors $\mathbf{f}$, called aspects in the EBA model. The probability of choosing option i over option j is given as

$$p_{ij} = \frac{\sum_k w_k f_{ik}(1 - f_{jk})}{\sum_k w_k f_{ik}(1 - f_{jk}) + \sum_k w_k f_{jk}(1 - f_{ik})}, \quad (1)$$

where f_{ik} denotes the k^{th} feature of option i and w_k is the positive weight associated with it. The greater the weight of a feature the heavier its influence on the choice probabilities. The sum $\sum_k w_k f_{ik}(1 - f_{jk})$ collects the weights for all the aspects that option i has but option j does not have. Therefore, the choice between two alternatives depends only on the features that are not shared between the two. In this way the EBA model can account for the effects of similarity on choice. If the options are characterized only by unique aspects, i.e. no option shares any feature with any other option, the BTL model is recovered.

If one option i has all the features that another alternative j has and more features on top of these then i will always be preferred over j. This is a reasonable assumption but in real choice data it can happen that subjects occasionally fail to choose alternative i because of error or lack of concentration (Kuss et al., 2005). In order to make our inference more robust to these lapses we add a lapse probability ε to the choice probabilities that we fix to a small value. Thus the choice probabilities become $\tilde{p}_{ij} = (1 - \varepsilon)p_{ij} + 0.5\varepsilon$.

Let us denote the number of times that i was chosen over j in a paired comparison experiment by x_{ij}. It is assumed that x_{ij} is binomially distributed

$$P(x_{ij}) = \begin{pmatrix} x_{ij} + x_{ji} \\ x_{ij} \end{pmatrix} (\tilde{p}_{ij})^{x_{ij}} (1 - \tilde{p}_{ij})^{x_{ji}}, \quad (2)$$

and is independent of all other comparisons in the experiment. We can then write the likelihood of all the observed choices in a paired comparison experiment as

$$P(X|F, \mathbf{w}) = \prod_{j=1}^{N} \prod_{i<j} P(x_{ij}|\mathbf{f}_i, \mathbf{f}_j, \mathbf{w}), \quad (3)$$

where X is a matrix that collects the results of all paired comparisons x_{ij}, F is a $N \times K$ binary matrix of features with entries f_{ik} for the k^{th} feature of option i, and $\mathbf{w}$ is a vector containing the weights w_k of all features.

To complete the model we need to specify the priors over the model parameters, i.e. the binary features and the weights. Since we do not know the features a priori we use a non-parametric prior over the binary feature matrix, the Indian buffet process (IBP). We put independent gamma [1] priors on the weights,

$$F \sim \text{IBP}(\alpha) \tag{4}$$

$$w_k \sim \mathcal{G}(1, 1). \tag{5}$$

IBP (Griffiths & Ghahramani, 2005) is a distribution over binary matrices with infinitely many columns with a parameter α that controls the sparsity of the matrix. Inspired by the derivation of the Chinese restaurant process by Pitman (2002), the process is described by imagining an Indian buffet offering an infinite number of dishes. Each customer entering the restaurant chooses the dishes that have been already sampled by other customers with probability proportional to their popularity. Then he also tries a number of new dishes dependent on the parameter α. The customers (rows of the matrix) are exchangeable and dishes (columns) are independent. For the EBA model the customers correspond to options in the choice set and the dishes correspond to features. Thus, the IBP prior implies that the ordering of the options is not important and the features are independent a priori.

The probability distribution defined by the IBP can also be derived by considering a finite feature matrix with K columns representing the features and taking the limit as $K \to \infty$, see Griffiths and Ghahramani (2005) for details. For the finite model, the conditional prior distribution for an entry f_{ik} in the feature matrix F is

$$P(f_{ik} = 1 | \mathbf{f}_{-i,k}) = \frac{m_{-i,k} + \alpha/K}{N + \alpha/K}, \tag{6}$$

where $\mathbf{f}_{-i,k}$ denotes the feature vector k with the ith element excluded, and $m_{-i,k}$ is the number of alternatives other than i that have feature k.

When we consider infinitely many features there might be some features that are shared between the options and some that are unique to an option. We will refer to both cases as the represented features. Furthermore, there will be infinitely many other features that no option has which we will refer to as the unrepresented features.

[1] $\mathcal{G}(\theta|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} \exp^{-\beta\theta}$

In the limit $K \to \infty$, the distribution of the features becomes

$$P(f_{ik} = 1 | \mathbf{f}_{-i,k}) = \frac{m_{-i,k}}{N}, \tag{7}$$

for $m_{-i,k} > 0$. For $m_{-i,k} = 0$, the probability

$$P(f_{ik} = 1 | \mathbf{f}_{-i,k}) = \frac{\alpha/K}{N + \alpha/K}, \tag{8}$$

approaches zero with $K \to \infty$. Considering infinitely many features results in a Poisson(α/N) distribution for the prior number of unique features for alternative i. Notice that $m_{-i,k} = 0$ for both the features that are unique to option i and the unrepresented features.

3. MCMC Inference

Inference for the above model can be done using MCMC techniques. We use approximate Gibbs sampling for updating the feature matrix F and the IBP parameter α and Metropolis Hastings updates for the weights $\mathbf{w}$. The sampling algorithm described has been summarized in Algorithm 1.

3.1. Feature Updates

Gibbs sampling for the feature updates requires the posterior of each f_{ik} conditioned on all other features $F_{-(ik)}$ and the weights $\mathbf{w}$. The conditional posterior for the represented features other than the features unique to the ith option can be obtained by combining the likelihood given in eq. (3) with the prior given in eq. (7),

$$\begin{aligned} &P(f_{ik} = 1 | X, F_{-(ik)}, \mathbf{w}) \\ &\qquad = \tfrac{m_{-i,k}}{Z} P(X | f_{ik} = 1, F_{-(ik)}, \mathbf{w}) \\ &P(f_{ik} = 0 | X, F_{-(ik)}, \mathbf{w}) \\ &\qquad = \tfrac{N - m_{-i,k}}{Z} P(X | f_{ik} = 0, F_{-(ik)}, \mathbf{w}), \end{aligned} \tag{9}$$

where Z is the normalizing constant.

The prior probability given in eq. (8) for $m_{-i,k} = 0$ approaches zero as $K \to \infty$. In the conjugate models, like the linear Gaussian model in Griffiths and Ghahramani (2005) or additive clustering in Navarro and Griffiths (2005), the parameters associated with the features can be integrated out. Therefore, the posterior distribution for the infinitely many features can be computed analytically. For the model we consider, the prior for the weights given in eq. (5) and the likelihood given in eq. (3) are not conjugate. The likelihood cannot be marginalized over the weights, hence the weights associated with each feature vector need to be represented explicitly. Therefore, we cannot compute the limiting posterior distribution over the infinite feature matrix.

We can obtain an approximation to the infinite case by truncating the number of Bernoulli trials with probability $\frac{\alpha/K}{N+\alpha/K}$ at a finite value K^*, and considering the joint posterior probability of these K^* features. Note the difference compared to a finite model where the number of total features is fixed at a certain value. For IBP there are always infinitely many features, most of them being unrepresented, hence they do not affect the likelihood. We use the truncation only as an approximation for the Gibbs sampling updates of the existing or new unique features. Therefore, the number of features the model can introduce is not bounded contrary to a model with a finite but large number of features.

The IBP models have close correspondences with the Dirichlet process (DP) models which allows infinite components in the mixture model. The unique features of an option can be thought of as the singleton components in the DP, and the unrepresented features as the mixture components that do not have any data associated with them. The sampling scheme we use for IBP is very similar to "Algorithm 8" of Neal (2000) for inference in the Dirichlet process models with non-conjugate priors.

We use K^* auxiliary variables to represent the possible values for the weights $\mathbf{w}^*$ of the features that are not associated with any other option. We associate weights of the existing unique features of option i with some of the auxiliary weights and draw values from the prior given in eq. (5) for the rest of the auxiliary weights. We denote the represented features that are not unique to option i as F_- and the K^* features that are not associated with any other option with F_l^*. All entries of F_l^* are zero except the ith row. There are 2^{K^*} possible combinations for the unique features for option i. We evaluate the posterior probabilities of all $l = 1, \ldots, 2^{K^*}$ possible F_l^* and sample from this distribution to decide on which to include. The joint posterior for F_l^* will consist of the Bernoulli probabilities of setting each feature in the ith row to 0 or 1 (with probability $\frac{\alpha/K^*}{N+\alpha/K^*}$) and the probability of the data given F_-, $\mathbf{w}_-$, F_l^* and $\mathbf{w}^*$,

$$\begin{aligned} &P(F_l^*|X, F_-, \mathbf{w}_-, \mathbf{w}^*) \\ &\qquad \propto P\left(F_l^*\right) P\left(X|F_l^*, F_-, \mathbf{w}_-, \mathbf{w}^*\right). \end{aligned} \tag{10}$$

3.2. Weight Updates

We update the weights using Metropolis Hastings sampling. We sample a new weight from a proposal distribution $Q(w_k'|w_k)$ and accept the new weight with probability

$$\min\left(1, \frac{P\big(w_k'|X,\, F,\, \mathbf{w}_{-k},\, w_k\, \lambda\big)}{P\big(w_k|X,\, F,\, \mathbf{w}_{-k},\, w_k',\, \lambda\big)} \frac{Q(w_k|w_k')}{Q(w_k'|w_k)}\right). \tag{11}$$

Algorithm 1 MCMC algorithm for EBA

Inputs: X, K^*
Initialize: F, W and α randomly
Repeatedly sample as follows:

for all objects $i = 1, \ldots, N$ **do** {Feature updates}
  for all represented features $k = 1, \ldots, K_+$ **do**
    if $m_{-i,k} > 0$ **then**
      update f_{ik} by eq. (9)
    else $\{m_{-i,k} = 0\}$
      set one of the auxiliary weights w_j^* to w_k
    end if
  end for
  sample values from eq. (5) for the w_j^* that are not yet assigned a value
  remove unique features of i from F to get F_-
  remove corresponding weights from $\mathbf{w}$ to get $\mathbf{w}_-$
  for all $l = 1, \ldots, 2^{K^*}$ possible F_l^* **do**
    calculate the posterior $P\left(F_l^*|X, F_-, \mathbf{w}_-, \mathbf{w}^*\right)$
  end for
  pick a feature combination F_l^* with probability proportional to its posterior
  update F to be the combination of F_- and F_l^*
  update $\mathbf{w}$ to be the combination of $\mathbf{w}_-$ and $\mathbf{w}^*$
  remove the zero columns from F and the corresponding weights
  discard F_l^* and $\mathbf{w}^*$
end for

for all weights $k = 1, \ldots, K_+$ **do** {Weight updates}
  draw a candidate w_k' from the proposal eq. (12)
  accept the new value with probability eq. (11)
end for
sample α from eq. (13)

As the proposal distribution we use a gamma distribution with mean equal to the current value of the weight, w_k, and standard deviation proportional to it,

$$Q(w_k'|w_k) = \mathcal{G}(\eta w_k, \eta/w_k). \tag{12}$$

We adjust η to have an acceptance rate around 0.5.

Note that there are infinitely many weights that are associated with the infinitely many features. Since the unrepresented features and their weights do not affect the likelihood, we need to only consider the weights that are associated with the represented features.

3.3. Update for IBP parameter α

The IBP parameter α affects the number of represented features therefore updating this parameter would give more flexibility to the model. We put a

Table 1. Choice probabilities for the Paris-Rome example. Columns are chosen over rows.

	P+	P	R	R+
P+	0.50	0	0.48	0.50
P	1	0.50	0.50	0.52
R	0.52	0.50	0.50	1
R+	0.50	0.48	0	0.50

vague gamma prior on α,

$$\alpha \sim \mathcal{G}(1, 1).$$

The likelihood for α can be derived from the joint distribution of the features given in Equation 34 of Griffiths and Ghahramani (2005),

$$p(F|\alpha) = \alpha^{K_+} \exp\big(-\alpha \sum_{j=1}^{N} \frac{1}{j}\big),$$

where K_+ is the number of represented components and N is the number of options. Combining this likelihood with the prior, we get the posterior distribution

$$p(\alpha|F) = \mathcal{G}\big(1 + K_+, 1 + \sum_{j=1}^{N} \frac{1}{j}\big). \tag{13}$$

4. Experiments

In this section, we present empirical results on an artificial data set and on real data. Both data sets have been considered in the choice model literature.

We observed some feature configurations in which the Markov chain got stuck due to the incremental updates of the Gibbs sampler and the likelihood function not being smooth. To avoid this problem, we used a two-part feature matrix model which is a combination of the infinite model and a finite model with one unique feature per alternative (equivalent to BTL), with corresponding weights. We kept the features of the finite part fixed while updating the weights for both finite and infinite parts so that if the evidence is in favor of an option not having a unique feature the weight for this feature can go to very small values. We report results obtained using this two-part model.

We truncate trials for the new features at $K^* = 5$ and set the lapse parameter to $\varepsilon = 0.01$. We initialize the parameters α, F and $\mathbf{w}$ randomly from their priors.

4.1. Paris-Rome

We first consider an example given by Tversky (1972). It was constructed as a simple example that the BTL model cannot deal with. We will use this example to illustrate that the EBA model with infinitely many latent features can recover latent structure from choice data.

Consider the choice between two trips that are on offer at a travel agency: one to Paris (P) and one to Rome (R). Another travel agency offers exactly the same trips there is an additional small bonus. to Paris and to Rome except that there is an additional small bonus. We denote these options by $P+$ and $R+$, respectively. The options in the choice set consist of P, $P+$, R, $R+$. We assume that the decision maker assigns the same value to the trip to Paris and to the trip to Rome. Hence, she will be equally likely to prefer the trip to either city. We denote the probability that Paris is chosen over Rome with $P(P, R) = 0.5$. As it is always better to get a bonus than to not get it we can assume that when given the options P and $P+$ she would choose $P+$ with certainty, i.e. $P(P+, P) = 1$. However, since a small bonus will not influence the choice between the two cities, $P(P+, R)$ will be close to 0.5, and likewise for $P(R+, P)$. We assume that the trip itself to either city has a feature with value w and the bonus is worth $0.01w$. Thus, the feature matrix is as shown top left of Figure 1. The alternatives P and $P+$ share the feature of being the trip to Paris and R and $R+$ share the feature of being the trip to Rome. $P+$ and $R+$ share the small bonus denoted by the $ column. Note that there might be some features that trips to either city possess such as taking the plane, going to Europe, etc. Since these features will be common to all options in the choice set they do not affect the choice probabilities. The matrix of choice probabilities calculated assuming the EBA model is shown in Table 1.

We generated choice data of 100 comparisons for each pair using these probabilities and used the EBA model with infinitely many latent features (iEBA) to infer the latent structure of the options. The results of a sample run for the iEBA model are shown in Figure 1. The top row shows the feature matrix that is used to generate the data, histogram of the frequencies for the represented features and the posterior distribution of the IBP parameter α. The plots below show the change in the log likelihood, the number of represented features of the IBP part of the feature matrix and α over the sampling iterations. It can be seen from the trace plots that the burn-in time is very short and the chain seems to mix well.

The iEBA model takes only the choice data and the truncation level as input, and it can recover the feature matrices given in Figure 1 from the data. Some

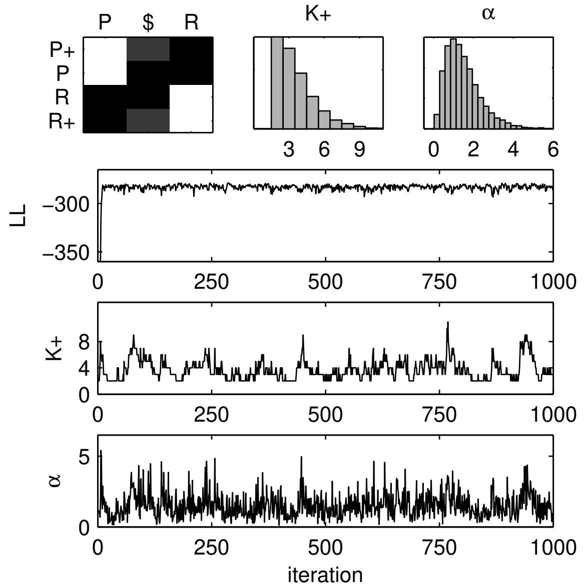

Figure 1. Feature matrix representation and simulation results on the toy data: the choice between trips to Paris and Rome. Top left: Features weighted by the associated values are shown. Rows correspond to the alternatives and columns correspond to the features. Darker means smaller in amplitude. The alternatives P and $P+$ share the feature of being the trip to Paris and R and $R+$ share the feature of being the trip to Rome. $P+$ and $R+$ share the small bonus denoted by the \$ column.

sample feature matrices from the chain are depicted in Figure 2 which shows that the model successfully finds the latent feature representation that was used to generate the data. Note that the mapping from the choice probabilities to features is not unique, that is, several latent feature representations may result in the same choice probabilities. The important point is that the model can infer that there are features that only P and $P+$ share, and there are features that only R and $R+$ share. The small bonus that $R+$ and $P+$ have in common is represented as two different features with similar weights. Models on a larger set of alternatives might result in feature representations that cannot be interpreted easily, as will be seen in the next example.

4.2. Celebrities

As a second example we analyze real choice data from an experiment by Rumelhart and Greeno (1971) that is known to exhibit patterns that the BTL model cannot capture. Subjects were presented with pairs of celebrities and asked "with whom they would prefer to spend an hour of conversation". There are nine celebrities that were chosen from three groups: three politicians, three athletes and three movie stars. Individuals within each group are assumed to be more

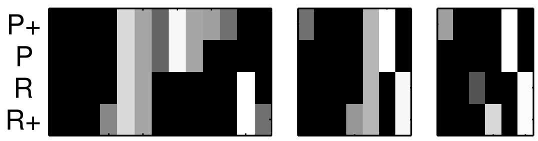

Figure 2. Random binary feature matrix samples from the chain for the Paris-Rome example. The gray level indicates the weight of each feature. Rows correspond to options and columns to features. Only the represented features are shown. It is inferred that there are features that only P and $P+$ share, and only R and $R+$ share. Note that there are also some features common to all options which do not affect the likelihood. The first four columns in these figures are the fixed features, remaining columns are from the posterior of the infinite part. Note that the weights of the fixed unique features for P and R are very close to zero, which is expected since they are not supported by the data. The unique features for $P+$ and $R+$ have small weights representing the small bonus.

similar to each other and this should have an effect on the choice probabilities beyond the variation that can be captured by the BTL model. The choice probabilities could be captured better by a feature matrix that has unique features for each individual plus one feature for being a politician, one feature for being an athlete and one feature for being a movie star. The assumed feature matrix is shown in Figure 3. This feature matrix can also be depicted as a tree therefore we refer to this model as the tree EBA model (tEBA) (Tversky & Sattath, 1979).

We modeled the choice data with different specifications for the EBA model: the model which assumes all options to have only unique features (BTL), the EBA model with tree structure (tEBA), two finite EBA models with the number of features fixed to be 12 and 15 (EBA12 and EBA15), and the EBA model with infinitely many latent features (iEBA). Although the experiment by Rumelhart and Greeno (1971) was

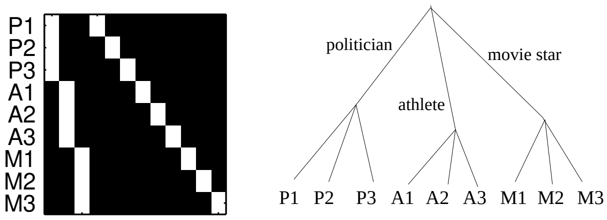

Figure 3. Representation of the assumed features for the celebrities data. The tree structure on the right shows nine celebrities of three professions. Each individual is assumed to have a unique feature and one feature that he shares with the other celebrities that have the same profession. The left panel shows this assumed feature matrix which is used for training the tEBA model.

designed with the tree structure depicted in Figure 3 in mind, we do not know the true latent features that lead to the choices of the subjects. We compare the predictive performance of each model using a a leave-one-out scenario.

We train the models on the data of all possible pairwise comparisons except one pair. We then predict the choice probability for the pair that was left out and repeat this procedure for all pairs. We evaluate the performance of each model by the negative log likelihood on the mean of the predictive distribution[2] for each pair. The negative log likelihood of each model averaged over the 36 pairs is shown in Table 2. For better comparison we also report the values for a baseline model that always predicts 0.5 and the upper bound that could be reached by predicting the empirical probabilities exactly. Furthermore, to see how much information we gain over the baseline model by using the different models we report an information score for each single paired comparison: The negative log likelihood averaged over the pairs and number of comparisons in bits with the baseline model subtracted.

The choice set was designed with the tree structure in mind. The iEBA and the tEBA models have the best predictive performance on average. This shows that the underlying structure could be successfully represented by the infinite model. However, we cannot observe the tree structure in the latent features that are found by the iEBA. Note that different feature representations can result in the same choice probabilities. The mean number of represented features for the iEBA model for different pairs is between 30 and 50—much more than the number of features in tEBA. This explains why the average performance of EBA12 and EBA15 is worse than that of iEBA and tEBA even though they could implement the tree structure in principle. As we cannot know how many features will be necessary beforehand this is a strong argument for using a non-parametric prior. As expected, the BTL model cannot capture as much information as the other models.

Figure 4 shows a more fine-grained analysis of the BTL, tEBA and iEBA models. Each point corresponds to one paired comparison: the negative log likelihood of one model versus the negative log likelihood of another model. Out of 36 pairs BTL has a smaller log likelihood for 23 of the pairs when compared to tEBA and 26 when compared to iEBA. However, it can be seen that the bad performance of the BTL model on average is also due to the fact that it cannot capture the probabilities of some pairs at all. The iEBA and tEBA likelihoods are comparable although there are some pairs on which iEBA performs better than tEBA, and vice versa.

Table 2. Predictive performance of different models: The baseline model that always predicts a probability of 0.5 (BASE), the Bradley-Terry-Luce model (BTL), the finite EBA model with 12 features (EBA12), finite EBA model with 15 features (EBA15), EBA model with tree structure (tEBA), the EBA model with IBP prior (iEBA), and for comparison the empirical probabilities (EMP). NLL: The negative log likelihood values on the mean predictive probabilities. IS: Information score in bits (the information gain compared to the model that always predicts 0.5). NLL and IS both express the same values in different scales.

MODEL	NLL	IS
BASE	17.57	0
BTL	4.66	0.0795
EBA12	4.50	0.0806
EBA15	4.31	0.0817
tEBA	3.95	0.0839
iEBA	3.92	0.0841
EMP	2.89	0.0905

5. Discussion

EBA is a choice model which has correspondences to several models in economics and psychology. The model assumes the choice probabilities to result from the non-shared features of the options. We have suggested to use an infinite latent feature matrix to represent unknown features of the options. The usefulness of the EBA model has been hampered by the lack of such a method. We showed empirically that the infinite model (iEBA) can capture the latent structure in the choice data as well as the handcrafted model (tEBA). For data for which we have less prior information it might not be possible to handcraft a reasonable feature matrix.

We have described a sampling algorithm for inference in the EBA model that could deal with the non-conjugacy in the prior of the weights. So far inference for the IBP has only been considered for conjugate priors on the parameters associated with the feature matrix. We have described a sampling algorithm for inference in the EBA model that could deal with the non-conjugacy in the prior of the weights. The main theoretical contribution of this work is the extension

[2]The loss function $\mathcal{L}(\hat{\theta}, \theta) = -\theta \log\hat{\theta} - (1-\theta)\log(1-\hat{\theta})$ expresses the discrepancy between the true probabilities θ and the predicted probabilities $\hat{\theta}$. The mean of the predicted probabilities minimizes the expected loss and therefore we take this as a point estimate.

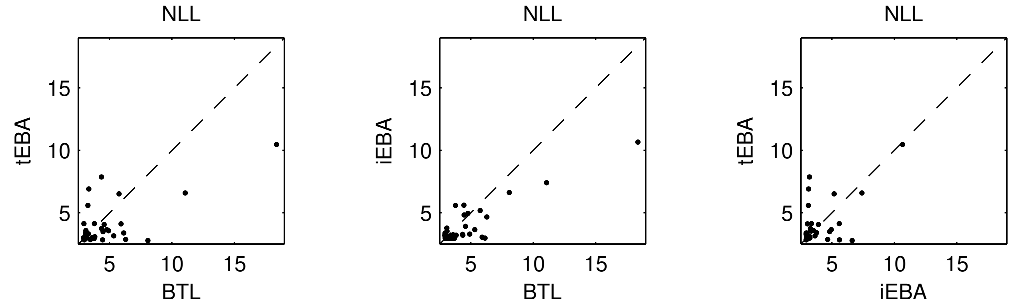

Figure 4. The negative log likelihood of one model versus another. Each point corresponds to one paired comparison.

of MCMC inference using the IBP prior in the non-conjugate case. This may widen the applicability of the IBP prior considerably.

Some issues about our sampling algorithm need to be addressed. We approximate the limit distribution of the IBP prior by truncation. The quality of approximation depends on the truncation level K^*. But large values of K^* are prohibitive due to the exponential increase in computation. We are currently working on improving the computational cost and on evaluating the quality of approximation.

Different feature matrices can result in the same choice probabilities and therefore are not distinguished by the model. For example, features that are shared by all options do not affect the likelihood. Furthermore, only the ratio of the weights affects the likelihood. What seems to be a non-identifiability problem is not an issue for sampling since we are interested in inferring the choice probabilities, not the "true" features.

On a more conceptual side the non-identifiability of the model makes the samples from the posterior hard to interpret. For the celebrities data one might have hoped to find feature matrices that correspond to a tree or at least find matrices with some other directly interpretable structure. However, we can use the posterior to predict future choices from past data, assess the similarity of the options and cluster or rank them.

Acknowledgments: CER is supported by German Research Foundation (DFG) through grant RA 1030/1.

References

Bradley, R., & Terry, M. (1952). The rank analysis of incomplete block designs. I. the method of paired comparisons. *Biometrika*, *39*, 324–345.

Griffiths, T. L., & Ghahramani, Z. (2005). *Infinite latent feature models and the Indian buffet process* (Technical Report 2005-01). Gatsby Computational Neuroscience Unit, University College London.

James, L. F., & Lau, J. W. (2004). Flexible choice modelling based on Bayesian nonparametric mixed multinomial logit choice models. *Submitted*.

Kuss, M., Jäkel, F., & Wichmann, F. A. (2005). Bayesian inference for psychometric functions. *Journal of Vision*, *5*, 478–492.

Luce, R. (1959). *Individual choice behavior*. New York: Wiley.

McFadden, D. (2000). Economic choice. In T. Persson (Ed.), *Nobel lectures, Economics 1996-2000*, 330–364. Singapore: World Scientific Publishing.

Navarro, D., & Griffiths, T. L. (2005). Bayesian additive clustering. *Proceedings of AML, 2*.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*, 249–265.

Pitman, J. (2002). Combinatorial stochastic processes. Notes for Ecole d'Eté Saint-Flour Summer School.

Restle, F. (1961). *Psychology of judgment and choice: A theoretical essay*. John Wiley & Sons.

Rumelhart, D., & Greeno, J. (1971). Similarity between stimuli: An experimental test of the Luce and Restle choice models. *Journal of Mathematical Psychology*, *8*, 370–381.

Train, K. (2003). *Discrete choice methods with simulation*. Cambridge University Press.

Tversky, A. (1972). Elimination by aspects: A theory of choice. *Psychological Review*, *79*, 281–299.

Tversky, A., & Sattath, S. (1979). Preference trees. *Psychological Review*, *86*, 542–573.

Wickelmaier, F., & Schmid, C. (2004). A Matlab function to estimate choice model parameters from paired comparison data. *Behavior Research Methods, Instruments, & Computers*, *36*, 29–40.