Parameter Clustering in a High-Dimensional Multinomial Choice Model

Didier Nibbering*

Econometric Institute, Tinbergen Institute, Erasmus University Rotterdam

February 14, 2017

Abstract

Parameter estimation in multinomial choice models becomes infeasible when the number of alternatives is large. As the parameters are alternative-specific, the number of parameters grows with the number of alternatives. Furthermore, in many applications the explanatory variables describe clusters which enter the model as large sets of dummies. This paper develops techniques for data-driven clustering over outcome categories and explanatory dummy parameters in a multinomial probit setting. A Dirichlet process mixture encourages parameters to cluster over the choice categories or explanatory categories, which favours a more parsimonious model without imposing any model restrictions. Simulation studies show that parameter clustering can improve greatly upon standard choice models in terms of predicting categories and fitting the category probabilities for many choice alternatives. The methods are applied to a high-dimensional choice data set of holiday destinations of a Dutch household panel.

Keywords: large choice sets, parameter clustering, multinomial probit model, high-dimensional models

JEL Classification: C11, C14, C25, C35, C51

*Address for correspondence: Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, e-mail: nibbering@ese.eur.nl

1 Introduction

Many multinomial choice problems involve large choice sets with hundreds of alternatives. Since the parameters in multinomial choice models are alternative specific, estimation methods become computationally infeasible when the choice set is very large. Different approaches for managing the number of parameter estimates are used in the literature. Researchers focus only on a subset of the alternatives, or the alternatives are arbitrarily aggregated to a higher level. This is clearly not a solution when all available alternatives are of interest. Moreover, considering subsets of alternatives lead to biased parameter estimates (Zanutto and Bradlow, 2006). As a solution, complex models decrease the number of parameters by introducing latent variables which explain the outcome categories (Ho and Chong, 2003; Jacobs et al., 2016). Instead of relating the explanatory variables directly to the different alternatives, they are related to a smaller set of latent variables.

In many applications with large choice sets, the number of parameters do not only increase in the number of outcome categories, but also in the number of explanatory variables. For instance, when a choice can be explained by past choice behavior, the choice in the previous time period enters the model as a large number of dummies constructed from the categorical explanatory variable consisting of past choice behavior.

This paper contributes to the literature of large choice sets by developing novel techniques for clustering over outcome categories and explanatory categories, to estimate the potentially high-dimensional model parameters in a multinomial choice model. By definition, the parameters in a multinomial choice model vary over the outcome categories. Instead of selecting subsets, or arbitrarily pooling outcome categories based on expert opinion, we propose a data-driven method for clustering outcome categories. A Dirichlet process mixture encourages parameters to cluster over outcome categories, which favours a more parsimonious model with efficient parameter estimates. We deal with a large number of categories in categorical explanatory variables by aggregating categories to a small set of dummies. Instead of pooling categories in dummy variables on expert opinion, we use a Dirichlet process mixture algorithm to pool in a data-driven way.

The multinomial choice problem is modelled by a multinomial probit model. Since the multinomial probit model can be presented by a set of normally dis-

tributed latent variables, this model can be extended by clustering algorithms developed for mixtures of normals. The Dirichlet process mixture sample algorithm of Ishwaran and James (2002) can be relatively straightforward applied to the multinomial probit setting for clustering over observations of individuals. However, the sample algorithm has to be modified and extended to suit clustering over outcome categories or categorical explanatory variables. Burda et al. (2008) models heterogeneity by a Dirichlet process mixture model on individual specific parameters. This paper serves as a good introduction to Dirichlet process mixture in a choice model. The Dirichlet process mixture allows for an infinite number of clusters but encourages clustering, which results in a parsimonious model. This means that we can estimate parameters in a large-dimensional choice model in several directions, without imposing any model restrictions. Moreover, the model incorporates the uncertainty in the number of clusters, which is ignored by fixing a priori the number of clusters or the distribution over the clusters, by jointly estimating the number of clusters, cluster assignments, and model parameters.

In this paper, we contribute to the literature in a number of directions. First, we propose a novel way of dealing with large choice sets without reducing the number of outcome categories. In practice, researchers make parameter estimation in choice models feasible by restricting their analysis to only a subset of the full choice set. When the smaller subset is obtained form selecting certain alternatives, parameter estimates are influenced by the selection mechanism (Zanutto and Bradlow, 2006). On the other hand, alternatives are aggregated to get a smaller choice set. However, Wasi et al. (2012) argues that using the full choice set is very informative. McFadden (1978) showed that using random subsets of the full choice set has no effect on the consistency of parameter estimates in the multinomial logit model. However, this result does not hold in the presence of unobserved individual hetergeneity (Wasi et al., 2012). By encouraging the clustering of outcome categories in a data-driven way, our approach is scalable to large choice sets, while abstaining from arbitrarily selecting or aggregating alternatives.

Second, we add to the literature of consideration sets by interpreting the clusters of outcome categories as different sets of preferences. Manzini and Mariotti (2014) argue that before making a final choice, individuals first form a consideration set from the total choice set, from which a final choice is made by maximizing the preference relation over the subset. Hauser (2014) discusses theoretic heuristics

and Liu and Arora (2011) several designs for econometric modelling this so called consider-then-choose model. Terui et al. (2011) and Chiang et al. (1998) find empirical evidence for consideration sets in a marketing context. Van Nierop et al. (2010) state that considering all possible consideration sets with a large number of alternatives is in practice infeasible. However, clustering outcome categories in different sets using Dirichlet process mixtures can be scaled up to large choice sets. Moreover, these sets can be interpreted as different preference sets. Mehta et al. (2003) explains consideration sets from limited information-processing capabilities, which limits the comparison of utilities to a subset of alternatives. This reasoning can also be used to explain preference sets. Individuals reduce the choice task by only making sets of alternatives between which they are indifferent, while there is a preference relation across the sets. In fact, a preference set is a generalization of the consideration set, since the probabilities of choosing one set can still be estimated equal to zero.

Third, we are, as for as we know, the first who overcome the problem of high-dimensional categorical explanatory variables by clustering over the explanatory categories. Since categorical explanatory variables are included in a model as a set of dummies, with for each category a dummy variable, parameter uncertainty already increases substantially by including only one high-dimensional categorical explanatory variable. Clustering over the categories boils down to aggregating dummies to a smaller set of dummies to gain in efficiency. The most common approach in the literature, is to cluster categories based on expert opinion. Evidently, this can lead to spurious results when the expert is wrong. On the other hand, all kinds of regularization techniques can be applied to the set of dummies, such as factor analysis or lasso (Naik et al., 2008). However, we explicitly exploit the categorical nature of the data to cluster the categories in the most efficient way in relation to the dependent variable.

We illustrate the practical implications of the clustering techniques in simulation studies and in an empirical application on holiday destinations of a Dutch household panel. We show that when parameters are clustered over outcome categories or explanatory categories in the data generating process, estimating the parameters in a standard choice model can lead to biased and noisy estimates, where a cluster model results in more precise and efficient posterior densities. The cluster model finds 12 different clusters out of the 66 categories of holiday

destinations and we show how this classification results in efficient estimates and interpretable results in a high-dimensional choice setting.

The outline of the remainder of this paper is as follows. Section 2 discusses the general specification of the multinomial probit model, introduces the clustering models, and explains parameter inference by Bayesian methods. Section 3 explains the properties of the clustering methods using simulated data and compares the performance to a standard multinomial choice model. Section 4 discusses the empirical application of the methods on holiday destinations of Dutch households. We conclude with a discussion in Section 5.

2 Methods

This section discusses the specification and parameter estimation of a high dimensional multinomial choice model. Section 2.1 introduces the baseline specification of a multinomial probit model. Section 2.2 specifies mixture models over the parameters in this model, and explains how parameter clustering over different dimensions makes parameter estimation feasible in high-dimensional settings. The clusters and parameter values are estimated by Bayesian methods. In Section 2.3, we specify the prior distributions and set up a Markov Chain Monte Carlo (MCMC) sampler. Moreover, we show how we can sample from the predictive density.

2.1 Multinomial Probit Model

Let y_i be an observable unordered outcome such that $y_i \in \{1, 2, \ldots, J\}$ and x_i be a K-dimensional vector with explanatory variables, where $i = 1, \ldots, N$ with N the number of individuals. Let $z_i = (z_{i1}, \ldots, z_{iJ})'$ be a $J \times 1$ vector of continuous unobservable latent variables, such that

$$y_i(z_i) = j \text{ if } z_{ij} = \max(z_i), \tag{1}$$

where $\max(z_i)$ is the largest element of the vector z_i. The latent variables are modeled as

$$z_i = \beta_i x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \Sigma), \tag{2}$$

where $\beta_i = (\beta_{i1}, \ldots, \beta_{iJ})'$ is a $J \times K$ matrix of coefficients, and ε_i an independent disturbance vector with covariance matrix Σ. We now have defined the conditional density $p(y_i|x_i, \beta_i, \Sigma)$, where the covariates in x_i are constant across different outcome categories, but the K-dimensional row vector of model parameters β_{ij} varies over outcome categories and individuals.

The parameters β_{ij} and Σ in the multinomial probit model specified in (1) and (2) are not identified (Bunch, 1991). First, $y_i(z_i + c) = y_i(z_i)$ for each scalar c. To overcome this additive redundancy we set $\beta_{i1} = 0$ for all i. However, we still have $y_i(cz_i) = y_i(z_i)$ for each positive scalar c. A conventional identifying assumption is to restrict the first element of the covariance matrix to be equal to one (McCulloch et al., 2000). Burgette and Nordheim (2012) restrict the trace of the covariance matrix to sample identified parameters. Instead of restricting the covariance matrix, McCulloch and Rossi (1994) specify in a Bayesian analysis of the multinomial probit model a proper prior for the unidentified model parameters and report the posterior of the model parameters up to a scaling. Imai and van Dyk (2005) introduce a new parameter to link identified to unidentified parameters.

Although these approaches lead to models with formally identified parameters, Keane (1992) shows that identification of multinomial probit models is extremely fragile because "it is difficult to disentangle covariance parameters from regressor coefficients". Many economic applications suffer from this problem, which is the reason that the multinomial probit model with a diagonal covariance matrix is most commonly used in applied research (Rossi et al., 2005). However, in practice, even in a diagonal covariance matrix different values for the variances can hardly be identified.

Beside the identification issues, the high-dimensional settings we study in this paper make the estimation of the large number of parameters in the (diagonal) covariance matrix infeasible. The number of parameters in the covariance matrix increases in the number of outcome categories J. Since we consider settings in which J grows large, we restrict the covariance matrix Σ in (2) to be the identity matrix. As a consequence of this restriction, the independence of irrelevant alternatives property holds.

Since the multinomial probit model with identity covariance matrix obeys the independence of irrelevant alternatives and the outcome probabilities are defined by intractable integrals, it seems convenient to use a logit instead of a probit

model. In contrast to the multinomial probit model, the outcome probabilities in the multinomial logit model have a closed form solution. However, the multinomial probit model can be represented by a set of normally distributed latent variables, where the distributions of the latent variables in the multinomial logit model are exotic or complex normal mixtures. Because of the Gaussian latent variables in (2), we can relatively straightforward apply clustering methods to the model parameters of the multinomial probit model.

2.2 Parameter Clustering

In high dimensional settings the number of model parameters in (2) easily grows large in various directions. First of all, when the number of individuals in the sample is large, it seems reasonable to assume that there is heterogeneity in the preferences for the different outcome categories. As Allenby and Rossi (1998) state, there is no reason to believe that these differences can solely be captured in the intercepts. Therefore, we allow for heterogeneity over the N individuals in all model parameters. Second, the number of possible outcome categories can be huge. Imagine, for example, the number of products we can choose from in a super market. Since the model parameters in the multinomial probit model are different for each outcome category j, a choice problem from many categories easily blows up the parameter space. Third, beside a dependent categorical variable, also a explanatory categorical variable can be included in the model. For instance, product choices in the current super market visit are related to purchases in previous visits. When we include all previous choice categories as dummies in the model, the number of parameters increase in the direction of K, the number of explanatory variables.

2.2.1 Clustering over Individuals

To allow for individual heterogeneity, we estimate clusters of individuals. Within a cluster individuals have the same parameter values, but parameter values are allowed to vary across clusters. We use a Dirichlet process prior to cluster over the individual specific parameters β_i. The Dirichlet process mixture model is defined

as,

$$z_i = \beta_i x_i + \varepsilon_i, \tag{3}$$

$$\varepsilon_i \sim \mathcal{N}(0, I_J), \tag{4}$$

$$\beta_i | P \sim P, \tag{5}$$

$$P | \alpha, H \sim DP(\alpha, H), \tag{6}$$

where the relation between the observable outcome y_i and the unobservable latent variable z_i is described in (1). The Dirichlet process is denoted by $DP(\alpha, H)$ with α a positive concentration parameter and H a continuous base distribution. The expectation over the Dirichlet process equals the base distribution and the concentration parameter governs the dispersion around the base distribution. When α large, the distributions P and H are more similar.

Sethuraman (1994) formulates the stick-breaking representation to show the construction of the Dirichlet process,

$$P = \sum_{l=1}^{\infty} p_l \delta_{\beta_l}, \tag{7}$$

$$p_l = V_l \prod_{l=1}^{l-1} (1 - V_l), \tag{8}$$

$$V_l \sim Beta(1, \alpha), \tag{9}$$

$$\beta_l \sim H. \tag{10}$$

where δ_l denotes a unit-mass measure concentrated at l, $\sum_{l=1}^{\infty} p_l = 1$, and we can also write $p = \{p_l\}_{l=1}^{\infty} \sim Stick(\alpha)$. This representation is formulated in terms of the clustering behavior of the individuals $1, \ldots, N$. Equivalently, we can describe the clustering behavior in terms of the classification variables $C_1, \ldots, C_N$,

$$z_i = \beta_{C_i} x_i + \varepsilon_i, \tag{11}$$

$$\varepsilon_i \sim \mathcal{N}(0, I_{J-1}), \tag{12}$$

$$C_i | p \sim \sum_{l=1}^{\infty} p_l \delta_l, \tag{13}$$

$$p \sim Stick(\alpha), \tag{14}$$

where β_{C_i} can vary over an infinite number of different clusters and the classification variables $C_i \in \{1, \ldots, L\}$ take integer values indicating the cluster for i,

identifying the β_l corresponding to a specific individual i. Thus, given the classification vector $C = (C_1, \ldots, C_N)$ one can describe the clustering behavior of individual i.

2.2.2 Clustering over Outcome Categories

To distinguish different outcome categories, the parameters in a multinomial choice model are different for each outcome category. Separate parameters for each outcome category can be estimated without additional restrictions as long as the number of parameters is smaller than the number of time periods. However, a large number of parameters increases parameter uncertainty. Moreover, hundreds of parameter estimates are not very insightful. Therefore, when subsets of states can be treated as a single state, this parsimonious model is to be preferred. Cramer and Ridder (1991) propose a statistical test, to test whether categories share the same parameter value. However, to test for all different combinations of subsets is computationally expensive and the order of tests can change the results. Therefore, researchers arbitrarily aggregate outcome categories into subsets in practice.

The Dirichlet process mixture model for clustering over individuals, proposed in Section 2.2.1, can also be applied to cluster over the outcome category dimension. Just as for a group of individuals, the Dirichlet process mixture infers whether a subset of categories can be treated as a single state, or whether its members are significantly different to distribute them over different clusters.

The Dirichlet process mixture model for clustering over outcome categories is defined as,

$$z_{ij} = \beta_{ij} x_i + \varepsilon_{ij}, \tag{15}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, 1), \tag{16}$$

$$\beta_{ij} | P \sim P, \tag{17}$$

$$P | \alpha, H \sim DP(\alpha, H), \tag{18}$$

for which we can derive the stick-breaking representation in the same way as for the cluster model over individuals in Subsection 2.2.1. So conditional on the

classification vector $C = (C_1, \ldots, C_J)$ we have,

$$z_{ij} = \beta_{iC_j} x_i + \varepsilon_{ij}, \tag{19}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0, 1), \tag{20}$$

$$C_j | p \sim \sum_{l=1}^{\infty} p_l \delta_l, \tag{21}$$

$$p \sim Stick(\alpha). \tag{22}$$

The parameters β_{ij} can be estimated by combining the clustering over the outcome categories with a cluster model over the individuals, or simply by imposing the restriction $\beta_{ij} = \beta_j$ for all $i = 1, \ldots, N$. Note that both allowing for individual and outcome category varying parameters can lead to identification problems.

Within a set of categories with the same parameter values, each category has the same probability of being chosen conditional on the explanatory variables. This means that individuals are indifferent between categories within the subset. However, they have a preference ranking across the clusters of outcome categories. This is related to the theory of consideration sets, in which individuals have a set of categories for which the probability to be chosen is zero, and form a preference relation over the remaining categories. The consideration set is defined as the categories for which the probabilities are nonzero. We generalize this consideration sets to preference sets, in which none of the preference sets is restricted to have categories with a probability of zero.

Instead of only considering a subset of categories, due to limited information-processing capabilities (Mehta et al., 2003), individuals can also form sets of alternatives between which they are indifferent. The choice task is now reduced to a preference relation over the different subsets, where in the consideration sets literature the preference relation is limited to the categories in the consideration set. Note that the framework of preference sets embeds consideration sets, since the probabilities of choosing a set can still be estimated equal to zero.

2.2.3 Clustering over Explanatory Categories

A categorical variable with a large number of categories can also be included as regressor in the model. For instance, there are no continuous regressors available but only categorical variables with many categories. Another example is that a

dependent variable consisting of a large number of categories can be explained in an autoregressive process. Including dummies for all different categories decreases efficiency or is even infeasible when number of categories is larger than the number of individuals in the sample. To make inference feasible we cluster over explanatory categories, using the same techniques as for clustering over individuals and outcome categories.

First we rewrite the model in (2) to make an explicit distinction in the regressor matrix $x_i = (w_i', d_i')'$, between the K_d dummies corresponding to the categories in the categorical explanatory variable, d_i, and the K_w remaining explanatory variables but without an intercept, w_i,

$$z_{ij} = \beta_{ij} x_i + \varepsilon_{ij} = \gamma_{ij} w_i + \kappa_{ij} d_i + \varepsilon_{ij}, \tag{23}$$

where $\beta_{ij} = (\gamma_{ij}', \kappa_{ij}')'$ and where the parameter values in $\kappa_{ij} = (\kappa_{ij1}, \ldots, \kappa_{ijK_d})$ correspond to the dummy categories in $d_i = (d_{i1}, \ldots, d_{iK_d})$.

The Dirichlet process mixture model for clustering over explanatory categories can now be defined as,

$$z_{ij} = \gamma_{ij} w_i + \sum_{k=1}^{K_d} \kappa_{ijk} d_{ik} + \varepsilon_{ij}, \tag{24}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0,1), \tag{25}$$

$$\beta_{ijk} | P \sim P, \tag{26}$$

$$P | \alpha, H \sim DP(\alpha, H), \tag{27}$$

for which we can derive the stick-breaking representation in the same way as for the cluster model over individuals in Subsection 2.2.1. So conditional on the classification vector $C = (C_1, \ldots, C_{K_d})$ we have,

$$z_{ij} = \gamma_{ij} w_i + \sum_{k=1}^{K_d} \kappa_{ijC_k} d_{ik} + \varepsilon_{ij}, \tag{28}$$

$$\varepsilon_{ij} \sim \mathcal{N}(0,1), \tag{29}$$

$$C_k | p \sim \sum_{l=1}^{\infty} p_l \delta_l, \tag{30}$$

$$p \sim Stick(\alpha). \tag{31}$$

The parameters β_{ijk} can be estimated by combining the clustering over the outcome categories with a cluster model over the individuals, or simply by imposing the restriction $\beta_{ij} = \beta_j$ for all $i = 1, \ldots, N$.

Within a cluster, dummies are aggregated to a new dummy variable, which results in a smaller number of dummies. As a result, the explanatory categories within one cluster, have the same effect on the dependent variable. As a consequence, the estimated relation between the categorical explanatory variable and the dependent variable can be spurious when the clustering is wrong. In contrast to pooling categories based on expert opinion or regularization techniques, the Dirichlet process mixture model clusters the categories in a data-driven way, in which also the relation to the dependent variable is taken into account.

2.3 Bayesian Inference

To estimate the parameters β_{ijk} we approximate the Dirichlet process mixture model by truncating the Dirichlet process at the Lth term by setting $V_L = 1$. For inference in the truncated Dirichlet process we build upon the blocked Gibbs sampler described by Ishwaran and James (2002) that is simpler than corresponding samplers for the full Dirichlet Process, while displaying favorable mixing properties. This sampler is extended to suit a multinomial choice model and we introduce new sampling steps for clustering over outcome and explanatory categories.

When we set the truncation level L equal to the number of available individuals, outcome, or explanatory categories, the truncated model is in practice equal to the full Dirichlet process mixture model. However, smaller values for L improve computational time significantly while results are effectively indistinguishable from results based on the full Dirichlet process when L is set to be much larger than the expected number of clusters in the data.

2.3.1 Prior Distributions

The cluster models are defined by a Dirichlet Process prior on the mixture distribution over different clusters for the coefficient vector. To complete the prior specification, we specify the base distribution H by the prior distribution for the

coefficients

$$\beta_{i1k}|c \sim \mathcal{N}(0,0) \text{ and } \beta_{ijk}|c \sim \mathcal{N}(0,c), \tag{32}$$

where $i = 1, \ldots, N$, $j = 2, \ldots, J$, $k = 1, \ldots, K$, and $c \in \mathcal{R}^+$. Moreover, we let the data determine the number of clusters by treating the concentration parameter α as unknown with a prior distribution,

$$\alpha|\eta_1, \eta_2 \sim \text{Gamma}(\eta_1, \eta_2), \tag{33}$$

where $\text{Gamma}(\eta_1, \eta_2)$ denotes a gamma distribution with mean η_1/η_2. The values $(\eta_1, \eta_2) \in \mathcal{R}^+$ directly effect the number of estimated clusters through the concentration parameter, where larger values for α encourage more distinct values for the coefficients.

2.3.2 Posterior Distributions

Ishwaran and Zarepour (2000) and Ishwaran and James (2002) derive a sample algorithm for finite normal mixture models using truncated Dirichlet process priors. Since the multinomial probit model can be represented by a set of Gaussian latent variables, as we show in (2), this algorithm needs only one extra sampling step for the latent variables to suit model parameter sampling in multinomial probit models.

The sample algorithm is developed for normal mixtures over the observations (y_i, x_i), which is relatively straightforward since clusters of observations are independent of each other. As a consequence, parameters corresponding to different clusters can be sampled independently. However, the parameters corresponding to one cluster of outcome categories are directly related to the parameters corresponding to another outcome category. The same holds for the explanatory categories; when the categorical regressor is correlated with another explanatory variable, there can be a high dependence between parameters corresponding to different category dummies.

Since the general setup of the Gibbs sampler for the three cluster dimensions is roughly the same, we first point out the sampling steps for clustering over individuals. We indicate which steps differ for different cluster dimensions, and the sample algorithm is followed by a detailed discussion of these steps.

Sample algorithm clustering over individuals Let $C^* = \{C_1^*, \ldots, C_m^*\}$ denote the current m unique values of C, and define the matrices $x = (x_1, \ldots, x_N)'$ and $z = (z_1, \ldots, z_N)$. Initialize the sampler by an initial draw for the latent variables from a standard normal distribution, and draw the concentration parameter from the prior distribution. Let the classification variables be equal to one, set the only nonzero weight $p_1 = 1$, and set the truncation level L. The sampling steps in each iteration of the sampler are:

Step 1. Sample the latent variables $z_{iy_i}|\beta_{iy_i}, y_i, x_i \sim \mathcal{N}_{+\max(z_{i,-j},0)}(\beta_{iy_i}x_i, 1)$ and sample $z_{ij}|\beta_{ij}, y_i, x_i \sim \mathcal{N}_{-\max(z_{i,-j},0)}(\beta_{ij}x_i, 1)$ for $j \neq y_i$ and $i = 1, \ldots, N$, where $z_{i,-j}$ denotes z_i without element z_{ij}, and $\mathcal{N}_+s$ and $\mathcal{N}_-s$ represent a normal distribution truncated from below and above by s, respectively.

Step 2. Sample the model parameters in β_l for $l \in C^*$ as

$$\beta_l|C, z, x \sim \mathcal{N}(b, B^{-1}), \qquad b = Z_l X_l B^{-1}, \qquad B = X_l{}' X_l + \frac{1}{c} I_{K_l}, \tag{34}$$

and for $l \in C - C^*$ as $\beta_l|C, z, x \sim \mathcal{N}(0, cI_{K_l})$. The construction of X_l, Z_l, and K_l differs per cluster dimension. For clustering over individuals we have the $r_l \times K$ matrix X_l with rows $x_{C_i=l}$, and Z_l is a $J \times r_l$ matrix with columns $z_{C_i=l}$. The number of values in C which equal l (the number of individuals in cluster l) is denoted by r_l and $K_l = k$.

Step 3. Sample the classification vector from $C|p, \beta, z, x$ with $\beta = (\beta_1, \ldots, \beta_L)$. This conditional distribution is different for each dimension over which we cluster. The specification of the conditional distribution of classification variables for clustering over individuals is

$$C_i|p, \beta, z_i, x_i \sim \sum_{l=1}^{L} \pi_{li}\delta_l, \tag{35}$$

for $i = 1, \ldots, N$, and where

$$(\pi_{1i}, \ldots, \pi_{Li}) \propto \left(p_1 \exp\left(-\frac{1}{2} \sum_{j=1}^{J} (z_{ij} - \beta_{1j}x_i) \right), \ldots, \right. \tag{36}$$

$$\left. p_L \exp\left(-\frac{1}{2} \sum_{j=1}^{J} (z_{ij} - \beta_{Lj}x_i) \right) \right). \tag{37}$$

Step 4. Sample the weights from $p|C, \alpha$ as

$$p_1 = V_1^*, p_l = (1 - V_1^*)(1 - V_2^*) \dots (1 - V_{l-1}^*)V_l^*, \text{ for } l = 2, \dots, L-1, \quad (38)$$

where

$$V_l^* \sim \text{Beta}\left(1 + r_l, \alpha + \sum_{k=l+1}^{L} r_k\right), \quad l = 1, \dots, L-1.$$

Step 5. Sample the concentration parameter

$$\alpha|p \sim \text{Gamma}\left(L + \eta_1 - 1, \eta_2 - \sum_{l}^{L-1} = 1 \log(1 - V_l^*)\right). \quad (39)$$

The first sampling step distinguishes the sampling algorithm for the multinomial probit model from a normal mixture model. From the five sampling steps, only step 2 and step 3 differ when clustering over other dimensions. The next two paragraphs discuss these steps in case of clustering over outcome categories or explanatory categories, respectively. For sake of simplicity, we suppress in these two paragraphs the subscript i of β and assume that β is constant over the individuals.

Sample algorithm clustering over outcome categories Since the $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{iJ})$, $i = 1, \dots, N$, in (2) are independently distributed, the cluster algorithm over individuals can sample β_{C_i} using all observations in cluster C_i, but independently from the observations assigned to other clusters than C_i. We us the same trick for sampling β_{C_j}, which differ over the different outcome categories instead of individuals. Since the covariance matrix Σ of ε_i in (2) equals the identity matrix, the ε_{ij} are also independent over $j = 1, \dots, J$. Therefore, x_i can be regressed on all information in z_{iC_j} for $i = 1, \dots, N$, where z_{iC_j} includes the latent variables corresponding to all outcome categories in the cluster C_j.

This means that Step 2 boils down to the sampling step in (34), where X_l and Z_l are specified as follows: Let r_l denote the number of outcome categories in cluster l. Define X_l as the $(r_l \times N) \times K$ matrix in which we stack r_l times the matrix x, and Z_l as the $1 \times (r_l \times N)$ matrix in which we stack the row vectors in z for which $z_{C_j=l}$. The classification variables are sampled similar as in Step

3: $C_j|p,\beta,z,x \sim \sum_{l=1}^{L} \pi_{lj}\delta_l$ for $j = 1,\ldots,J$, where here we do not sum over the outcome category dimension in the likelihood, but over the individuals,

$$(\pi_{1j},\ldots,\pi_{Lj}) \propto \left(p_1 \exp\left(-\frac{1}{2}\sum_{i}^{N}(z_{ij} - \beta_l x_i) \right), \ldots, \right. \tag{40}$$

$$\left. p_L \exp\left(-\frac{1}{2}\sum_{i}^{N}(z_{ij} - \beta_L x_i) \right) \right). \tag{41}$$

Sample algorithm clustering over explanatory categories To sample the classification variables in clusters over individuals or outcome categories, we exploit the conditional independence between the data in the different clusters. We sample the model parameters separately for each cluster, as in Step 2 of the sampling algorithm for clustering over individuals. Unfortunately, we cannot assume that clusters over different explanatory variables are independent. Therefore, the parameters of clusters of explanatory variables are sampled simultaneously.

These coefficients in the $J \times K_d$ matrix $\kappa = (\kappa_{1:J,C_1},\ldots,\kappa_{1:J,C_{K_d}})$ are sampled as for β_l in (34), where $X_l = X$ and $Z_l = Z$ are specified as follows: The dummy variables corresponding to the categories in cluster l are added up to create a new dummy indicating the cluster, $\tilde{d}_{il} = \sum_{k:C_k=l} d_{ik}$, which is the lth column of the row vector $X_i = (\sum_{k:C_k=1} d_{ik},\ldots,\sum_{k:C_k=m} d_{ik})'$. These row vectors are stacked in the $N \times m$ matrix $X = (X_1,\ldots,X_N)'$, $Z = (z_1 - \gamma w_1,\ldots,z_N - \gamma w_N)$, where $\gamma = (\gamma_1,\ldots,\gamma_J)'$, and $K_l = m$.

Since clusters of different explanatory variables are not necessarily independent, we cannot distinguish between likelihood contributions of each explanatory category, as we do for individuals or outcome categories. To measure likelihood contribution of each cluster value for the different category dummy coefficients, we introduce $\tilde{\kappa}_{jkl} = (\kappa_{C_1},\ldots,\kappa_{C_{k-1}},\kappa_1,\kappa_{C_{k+1}},\ldots,\kappa_{C_{K_d}})$, which is the coefficient vector κ_j based on the classification vector of the previous iteration of the sampler, where the coefficient corresponding to the kth dummy is replaced by the coefficient value of cluster l. Now the classification variables $C_k|p,\beta,z,x \sim \sum_{l=1}^{L} \pi_{lk}\delta_l$

for $k = 1, \ldots, K_d$, where

$$(\pi_{1k}, \ldots, \pi_{Lk}) \propto \left(p_1 \exp\left(-\frac{1}{2} \sum_i^N \sum_j^J (z_{ij} - \gamma_j w_i - \tilde{\kappa}_{jk1} d_i) \right), \ldots, \right. \tag{42}$$

$$\left. p_L \exp\left(-\frac{1}{2} \sum_i^N \sum_j^J (z_{ij} - \gamma_j w_i - \tilde{\kappa}_{jkL} d_i) \right) \right). \tag{43}$$

To sample the coefficients corresponding to the remaining explanatory variables in w_i, we add an extra sampling step to the sampling algorithm. Step 6 involves sampling of γ using the steps in (34) with $X_l = (w_1, \ldots, w_N)'$, $Z_l = (z_1 - \sum_{k=1}^{K_d} \kappa_{1:J,C_k} d_{1k}, \ldots, z_N - \sum_{k=1}^{K_d} \kappa_{1:J,C_k} d_{Nk})$, and $K_l = K_w$,.

2.3.3 Predictive Distributions

To construct a predictive density, we make use of the in-sample posterior clustering and the truncation level of the Dirichlet process. We draw the cluster assignment of an out-of-sample observation from the posterior mixture distribution of that observation. Moreover, since the truncation level is assumed to be much larger than the number of in-sample clusters, the model allows for new clusters of model parameters out-of-sample.

To draw from the predictive density of y_i, we draw the cluster assignment of this out-of-sample observation from the posterior mixture distribution. This distribution requires a different information set for each cluster dimension. When clustering over individuals, we need to observe a choice in the past of an individual, y_{iT}, to assign individual i to a cluster. Here we introduce the subscript $t = T$ to indicate the last available time period t for which we observe a choice for individual i. Therefore, individual i need to be observed in-sample to draw values for y_{iT+h} with $h = 1, 2, \ldots$. For clustering over outcome or explanatory categories, the posterior mixture distribution does not depend on individual characteristics. This means that after observing choices y_{iT} for $i = 1, \ldots, N$, we can draw from the predictive density of $y_{N+1,T+h}$, for $h = 0, 1, 2, \ldots$.

We simulate the predictive densities of y_{iT+h} for different individuals i and horizons h by using (1) and (2) in each iteration of the sampler, together with the parameter draws obtained in that sample iteration. In iteration (s) of the sampler,

we have

$$z_{iT+h}^{(s)} = \beta_{C^{(s)}}^{(s)} x_{iT} + \varepsilon_{iT+h}^{(s)}, \quad \varepsilon_{iT+h}^{(s)} \sim \mathcal{N}(0, I_J), \tag{44}$$

where we sample $C^{(s)}|\pi^{(s)} \sim \sum_{l=1}^{L} \pi_l^{(s)} \delta_l$, with $C^{(s)}$ and $\pi_l^{(s)}$ 1-dimensional vectors in case of clustering of individuals, and J- or K_d-dimensional vectors in case of clustering over outcome or explanatory categories, respectively. To obtain a draw from the predictive density of y_{iT+h}, use $z_{iT+h}^{(s)}$ as follows,

$$y_{iT+h}^{(s)}(z_{iT+h}^{(s)}) = j \text{ if } z_{ij,T+h}^{(s)} = \max(z_{iT+h}^{(s)}). \tag{45}$$

The draw from the predictive density is conditional on the draws for the model parameters $\beta_{C^{(s)}}^{(s)}$ and the conditional weights $\pi_l^{(s)}$ in iteration (s) of the sampler. Since the truncation level L is assumed to be much larger than the expected number of clusters in the data, future model parameter values can be drawn from new clusters which are not present in-sample.

3 Simulation Studies

We examine the practical implications of the developed cluster methods in three simulation studies. Each study points out the usefulness and effectiveness of a clustering method, by comparing the posterior results of the cluster algorithm to the results of a standard multinomial probit model.

The general set-up is the same in each simulation study. The data generating process takes the form

$$y_i(z_i) = j \text{ if } z_{ij} = \max(z_i) \tag{46}$$

$$z_i = \beta_i x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_J), \quad i = 1, \ldots, N, \tag{47}$$

with $N = 1000$. We sample values for the model parameters based on N generated observations, and use another N observations for out-of-sample analysis. Posterior results in each simulation study are based on 20,000 iterations of the Gibbs sampler, from which the first 10,000 are discarded. Visual inspection shows that this number of iterations is enough for convergence. The number of possible states is truncated at $L = 10$, and the prior parameter values are $c = 1$, $\eta_1 = 10$, and $\eta_2 = 1$.

We evaluate the in-sample and out-of-sample performance of the methods using the hitrate (HR); the fraction of correct predictions, defined as

$$HR = \frac{1}{NS} \sum_{s=1}^{S} \sum_{i=1}^{N} I(y_i = y_i^{(s)}) \tag{48}$$

where S denotes the number of samples from the predictive density, and $I(A)$ is an indicator function that equals one if event A occurs and zero otherwise. The hitrate only weights observations for which the predictions are right, but does not reward predictions which are close to the realized value. Therefore, we also evaluate model performance based on the implied category probabilities. The root mean squared error (RMSE) measure benefits posterior conditional category probabilities which are close to these probabilities implied by the data generating process,

$$RMSE = \sqrt{\frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} \Big(\widehat{P(y_i = j|x_i)} - P(y_i = j|x_i) \Big)}, \tag{49}$$

where $\widehat{P(y_i = j|x_i)} = \frac{1}{S} I(y_i^{(s)} = j)$, and $P(y_i = j|x_i)$ is simulated from the data generating process.

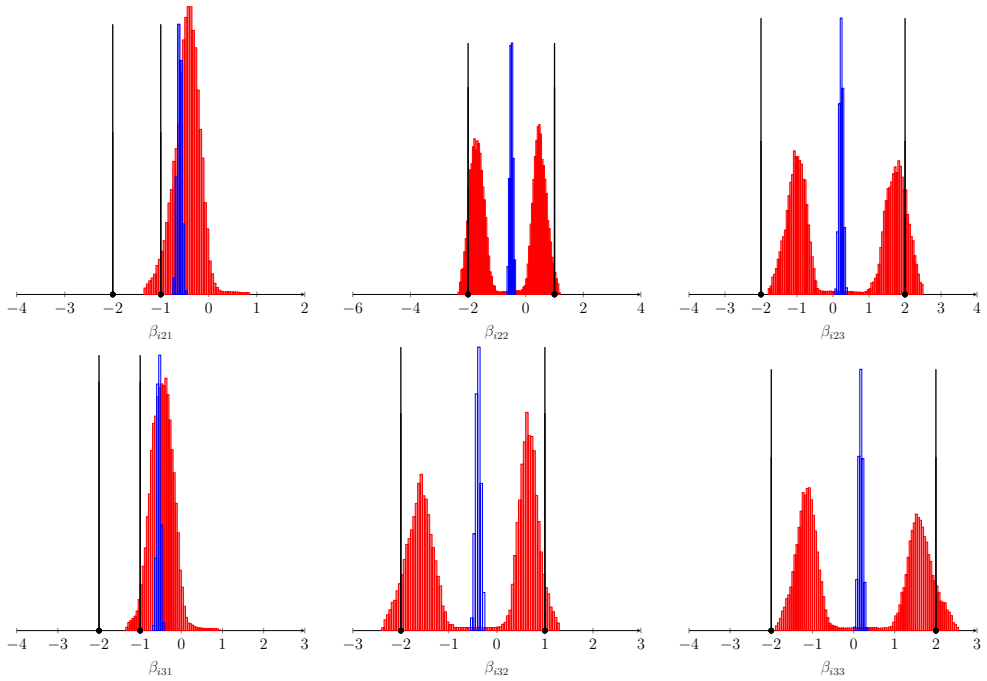
3.1 Clustering over Individuals

To show the effectiveness of the method in capturing unobserved heterogeneity, we generate a data set with two groups of individuals. Within these groups the effects of the explanatory variable on the outcome category probabilities are the same, but they are different across the two groups. The number of explanatory variables and outcome categories are relatively small to focus on the effect of clusters of individuals on the posterior results of a model that takes the heterogeneity into account compared to a model that assumes homogeneous individuals.

The vector x_i includes an intercept and two variables generated by the normal distribution,

$$(x_{i2}, x_{i3}) \sim \mathcal{N}(0, \Sigma_x), \quad \Sigma_x = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}, \tag{50}$$

Figure 1: Posterior Parameter Densities Clustering over Individuals



This figure shows posterior parameter densities of the individual cluster model (red) and a standard multinomial probit model (blue). The parameter densities are showed in the same order as in (51) and the black lines represent the parameter values in the data generating process.

so the number of explanatory variables equals $K = 3$, and we set the number of outcome categories $J = 3$. The model parameter values of the two groups equal,

$$\beta_{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ -2 & 1 & -2 \\ -1 & -2 & 2 \end{bmatrix}, \quad \beta_{(2)} = \begin{bmatrix} 0 & 0 & 0 \\ -1 & -2 & 2 \\ -2 & 1 & -2 \end{bmatrix}, \tag{51}$$

and an individual i is randomly assigned to one of the two groups, $P(\beta_i = \beta_{(1)}) = P(\beta_i = \beta_{(2)}) = 0.5$.

Figure 1 shows that the cluster model captures the individual heterogeneity, where the standard model finds parameter values close to zero. The parameter densities of the cluster model are bimodular, with modes corresponding to the two groups of individuals. The densities of the standard model have only one mode which is, for most parameters, located between the two modes of the cluster densities. As a result, there is almost no probability mass of the posterior densities

of the standard model at the true parameter values in the data generating process. As an exception, both models are not able to distinguish the different intercepts corresponding to the two groups of individuals. Although the standard model does not find the true parameter values, the variances of the densities are much smaller than for the cluster model. This can be explained by the fact that the cluster model does not only incorporate parameter uncertainty in the posterior densities, but also the uncertainty that comes along with estimating the number of different clusters and the cluster assignment.

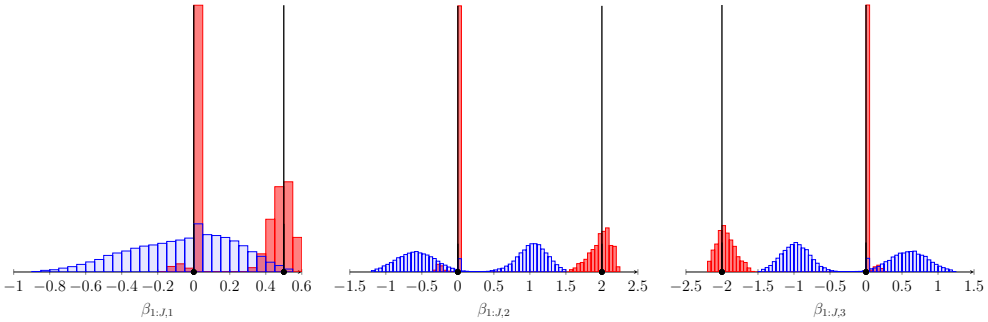
Table 1 shows that clustering over individuals improves hit rates and root mean squared errors upon a standard multinomial probit model, both in-sample and out-of-sample. We even find that the standard multinomial probit model is on both metrics outperformed by a naive method, which calculates category probabilities as percentage observed in the data and always chooses the category with the largest probability. This means that ignoring unobserved heterogeneity can lead to posterior results that are worse than simply counting the observed choices in the data. The clustering method accounts for unobserved heterogeneity, which results in larger hit rates and lower root means squared errors relative to the naive and standard methods. In other words, the method is more often right in predicting the outcome category and the posterior category probabilities are much closer to the probabilities implied by the data generating process.

3.2 Clustering over Outcome Categories

In contrast to the previous simulation study, we set the number of outcome categories to be large, and assume that there is only one homogeneous group of individuals. The parameter values corresponding to the different outcome categories are assigned to two groups. Within each group the relation between the explanatory variables and the category probabilities is the same, but the effects differ across the clusters of outcome categories.

We generate the explanatory variables in the same way as in (50). The number of outcome categories is large, $J = 50$, and since there is only one homogeneous group of individuals we suppress the subscript i in β_{ij}. The outcome categories

Figure 2: Posterior Densities Clustering over Outcome Categories



This figure shows posterior parameter densities of the outcome category cluster model (red) and a standard multinomial probit model (blue). The parameter densities are showed in the same order as in (52) and the black lines represent the parameter values in the data generating process.

are clustered into two groups, with parameter values

$$\beta_{(1)} = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}, \quad \beta_{(2)} = \begin{bmatrix} \frac{1}{2} & 2 & -2 \end{bmatrix}, \tag{52}$$

where $\beta_j = \beta_{(1)}$ for $j = 1, \ldots, 25$ and $\beta_j = \beta_{(2)}$ for $j = 26, \ldots, 50$.

Figure 2 shows that the cluster model is much more efficient in estimating parameters, with smaller variance in the posterior parameter densities corresponding to the different outcome categories compared to a standard multinomial probit model. In contrast to clustering over individuals, the clustering over outcome categories results in wider posterior densities for the standard model. The standard model estimates parameters for each outcome category separately. Since the outcome categories are clustered in the cluster model, less parameters need to be estimated which decreases the parameter uncertainty.

Table 1 shows that clustering over outcome categories improves the out-of-sample hit rate and the root mean squared errors upon a standard multinomial probit model. In-sample, the standard model achieves a slightly better hit rate, probably due the larger number of parameters. However, this gain in precision comes at the cost of efficiency. Therefore, the out-of-sample hit rate is best for the cluster model. Based on the RMSE there is no doubt which method performs best in estimating the category probabilities. The cluster method outperforms the standard multinomial probit model and the naive data method by a large margin.

3.3 Clustering over Explanatory Categories

The last simulation study considers groups of explanatory categories with similar effects on the outcome category probabilities. We include a categorical explanatory variable with a large number of categories in the regressor matrix, and assume that there are two groups of categorical dummies with different effects on the outcome category probabilities. Within in each group, the effect of each category is the same, while it differs across the groups.

The regressor matrix consists of a categorical variable d_i with a large number of categories $K_d = 50$, and a K_w continuous variables in w_i,

$$x_i = (w_i, d_i), \quad w_i = \frac{1}{2}\tilde{w}_i + \tilde{d}_i, \quad d_i = I_{K_d} \otimes 1_{N/K_d \times 1}, \tag{53}$$

where $\tilde{w}_i$ is drawn from a standard normal distribution and $\tilde{d}_i$ denotes the sum over the cumulative sum of the elements in the K_d-dimensional vector d_i divided by K_d. This construction ensures that the regressor w_i is correlated with the categorical explanatory variable d_i. In d_i, the observations are equally distributed over the different categories. To focus on the large number of explanatory categories, we consider only $J = 2$ different outcome categories. The model parameter values are

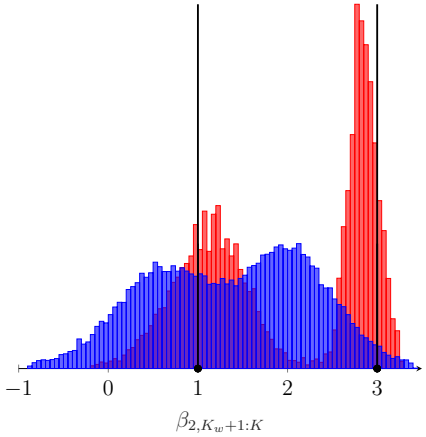
$$\beta_{1:J,1} = \begin{bmatrix} 0 \\ -2 \end{bmatrix}, \quad \beta_{1:J,(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \beta_{1:J,(2)} = \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \tag{54}$$

where $\beta_{1:J,k} = \beta_{1:J,(1)}$ for $k = 2, \ldots, 26$ and $\beta_{1:J,k} = \beta_{1:J-1,(2)}$ for $k = 27, \ldots, 51$.

Figure 3 shows that clustering over explanatory categories gains in efficiency relative to the standard multinomial probit model. The poster parameter densities of the standard model are much wider, indicating that the variance of the paramter draws is high. Similar to clustering outcome categories, clustering explanatory categories results in less parameters to be estimated, which explains the decrease in uncertainty in the densities corresponding to the cluster method.

Table 1 shows that clustering not only leads to more efficient posterior parameter densities, but also to higher hit rates and smaller root mean squared errors. Both in-sample and out-of-sample the standard multinomial probit model and the naive data method are outperformed.

Figure 3: Posterior Densities Clustering over Explanatory Categories



This figure shows posterior parameter densities of the explanatory category cluster model (red) and a standard multinomial probit model (blue). The parameter densities are showed in the same order as in (54) and the black lines represent the parameter values in the data generating process.

4 Empirical Application

We apply the clustering methods to survey data from a Dutch market research company. The data set consists of individual characteristics and details of all the reported holidays undertaken in 2015 by 5275 respondents. Among other things, respondents were asked to which country or region they have been for holidays. We select the destination country of the longest holiday in 2015 of each respondent as dependent variable. Since we only consider respondents who went abroad for their longest holiday, the number of individuals in the sample is 3527. We end up with 66 outcome categories and France is set as the base category. The explanatory variables include an intercept, the age of the respondent, the income of the respondent, calculated as the log mean of the income category, and dummies indicating whether the respondent lives in a single household or lives together with children. Appendix A shows descriptive statistics of all explanatory variables and Appendix B shows frequency counts for the outcome categories.

We estimate the parameters in the multinomial probit model defined in (1) and (2). To decrease parameter uncertainty and obtain interpretable results, we cluster over the outcome categories as explained in Section 2.2.2, and the parameters are

Table 1: In- and Out-of-Sample Performance Measures Cluster Methods

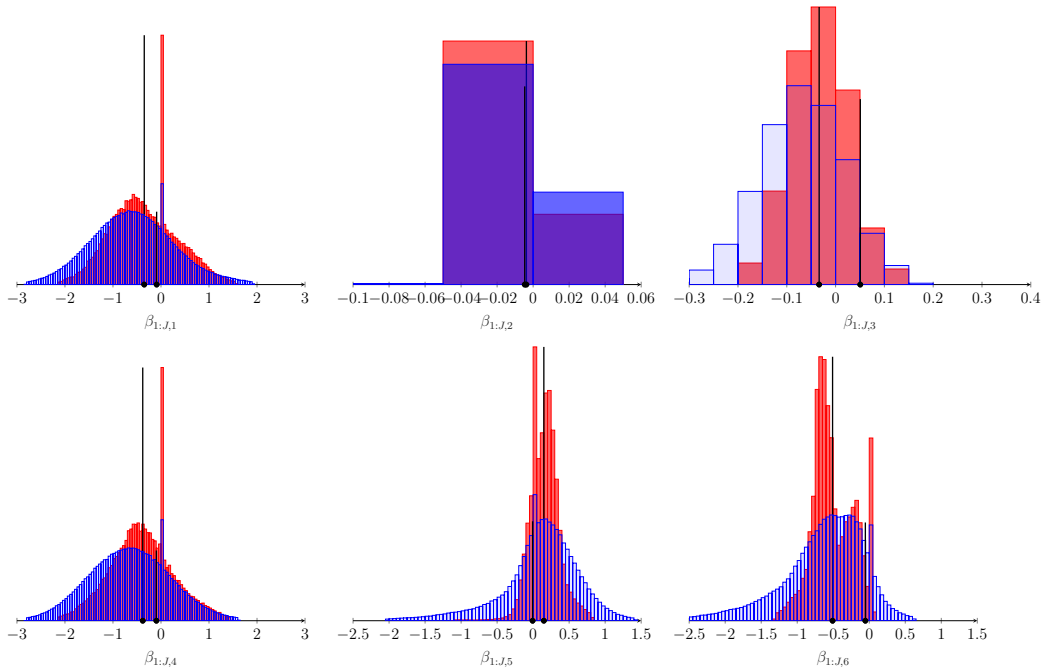
	cluster N		cluster J		cluster K_d	
method	in	out	in	out	in	out
			Hit Rates			
naive	0.493	0.480	0.030	0.036	0.646	0.660
standard	0.350	0.349	0.036	0.033	0.711	0.694
cluster	0.541	0.491	0.034	0.034	0.736	0.726
			RMSE			
naive	0.280	0.281	0.017	0.017	0.323	0.317
standard	0.293	0.293	0.010	0.010	0.156	0.154
cluster	0.180	0.180	0.005	0.005	0.129	0.129

This table shows the in-sample and out-of sample performance of different cluster methods measured by hit rates (48) and the root mean squared error (49). The performance of the clustering methods is compared to a standard multinomial probit model and a naive method in which the category probabilities are calculated as percentage observed in the data, and the category with the largest probability is always chosen.

sampled as outlined in Section 2.3. The prior parameters are set equal to $c = 1$, $\eta_1 = 2$, and $\eta_2 = 2$. The number of possible clusters is truncated at $L = 20$. The posterior results are based on 20,000 iterations of the sampler, from which the first 10,000 are discarded. Visual inspection shows that this number of iterations is sufficient for convergence.

Figure 4 shows the posterior parameter densities of the explanatory variables for all holiday destinations. Appendix B shows the posterior mean of each parameter density for each outcome category for the cluster model. After convergence, the sampler of the cluster model samples twelve different clusters of outcome categories in each iteration of the sampler. According to the posterior densities for the intercept, households have a lot of variation in base preferences over the different destinations, with values ranging from approximately -2 to 2. We find small effects for the age of the respondent and the gross annual income of the household.

Figure 4: Posterior Densities Clustering over Holiday Destinations



This figure shows posterior parameter densities of the explanatory variables in the outcome category cluster model (red) and a standard multinomial probit model (blue) for all holiday destinations. The parameter densities correspond to the intercept ($\beta_{1:J,1}$), age ($\beta_{1:J,2}$), income ($\beta_{1:J,3}$), no response for income ($\beta_{1:J,4}$), single household ($\beta_{1:J,5}$), and household with children ($\beta_{1:J,6}$). The black lines represent the posterior means for the parameters.

However, there is a broad range of different parameter values for the effects of being a single household or living in a household with children, for different holiday destinations.

Comparing the posterior densities of the standard multinomial probit model with the cluster model over outcome categories, we find the latter model to be more efficient. Although the cluster model accounts for the uncertainty about the number of clusters and the cluster assignments together with parameter uncertainty, sampling separate parameter values for each destination in the standard model results in much more noise. The posterior means over all outcome categories of the standard and cluster models are very close to each other. However, the shape of the posterior densities show, except for width, also other differences. Since the cluster model clusters the base category destination France with other

destinations, more probability mass is allocated to zero. The posterior densities of the standard model approximate the normal distribution for the intercept, income, and age effects. The densities for single households and households with children are skewed. Due to clustering of effects for different holiday destinations, the cluster model has multiple modes for these two dummies.

Figure 5 shows how the world is clustered based on the cluster model. The sets of destinations with the same color have the same probability of being chosen by a household. The households have a preference ranking across countries with different colors, but are indifferent between countries with the same color. To illustrate, households have the same probability of going to Italy or Spain, but the probability differs between Canada and the United States. The estimated clustering is very different from an ad hoc grouping based on, for instance, geographical location. Only in Europe, we already find nine different clusters. Popular destinations within Europe share clusters, such as Germany, Spain, Italy, and Greece. The red cluster includes other popular destinations of Dutch travelers; France and Turkey. There is one large cluster which includes a large part of Africa, Eastern Europe and Mexico.

5 Conclusion

In this paper we propose novel methods for parameter estimation in high dimensional choice models. To make parameter estimation feasible and gain in efficiency, we cluster parameters over outcome categories and explanatory categories. Clusters of outcome categories have an interpretation similar to consideration sets and clustering explanatory categories allows to account for categorical explanatory variables with a large number of categories. The methods build upon Dirichlet process mixture models applied to a multinomial probit model. We develop a Gibbs sampler for the multinomial probit model which can deal with the clustering models over different dimensions than individual observations. We find on simulated data excellent performance of the cluster methods compared to a standard choice model and an empirical application shows the practical implications on holiday destinations of a Dutch panel.

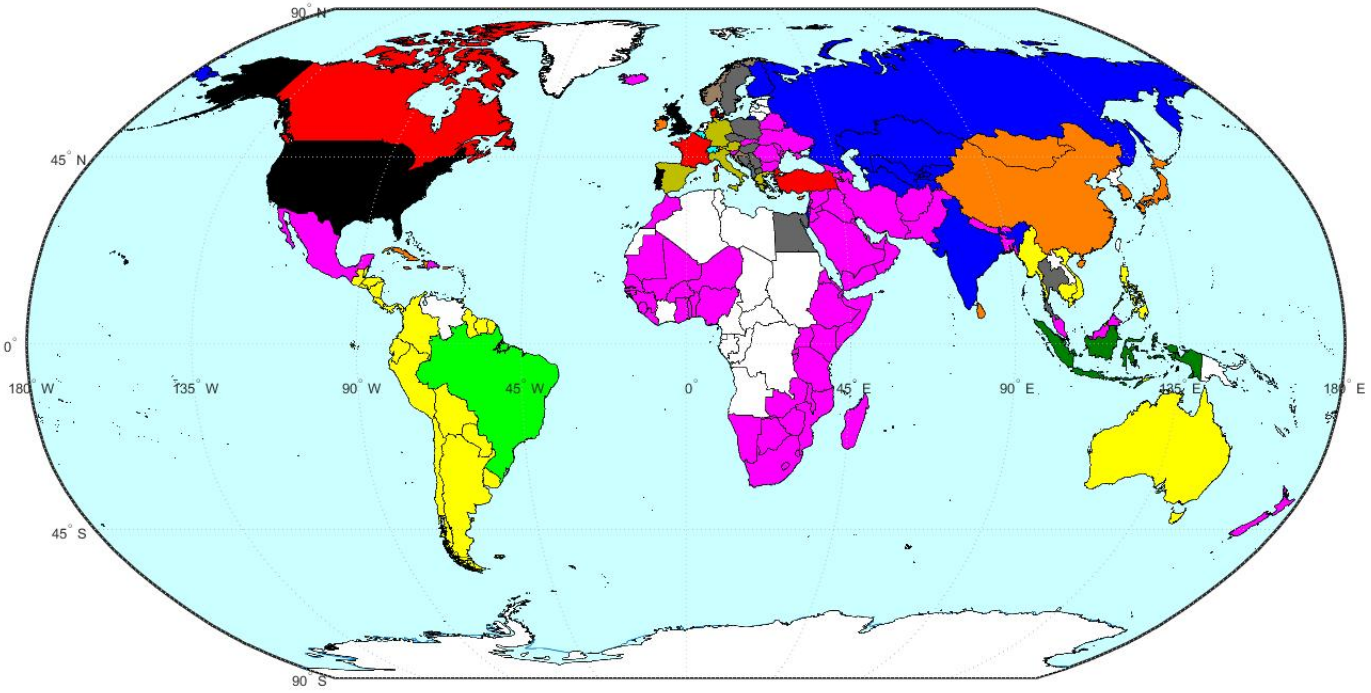


Figure 5: Cluster assignment of holiday destinations. Outcome categories with the same color are in the same cluster, where outcome categories are assigned to clusters with the highest posterior probability (we do not have observations about white regions). Appendix B shows which holiday destinations are in which outcome category. Note that this map shows a classification which is only very likely based on the results. Although the sampler seem to be stable after convergence, label-switching is theoretically possible.

References

Allenby, G. M. and Rossi, P. E. Marketing models of consumer heterogeneity. *Journal of econometrics*, 89(1):57–78, 1998.

Bunch, D. S. Estimability in the multinomial probit model. *Transportation Research Part B: Methodological*, 25(1):1–12, 1991.

Burda, M., Harding, M., and Hausman, J. A bayesian mixed logit–probit model for multinomial choice. *Journal of econometrics*, 147(2):232–246, 2008.

Burgette, L. F. and Nordheim, E. V. The trace restriction: An alternative identification strategy for the bayesian multinomial probit model. *Journal of Business & Economic Statistics*, 30(3):404–410, 2012.

Chiang, J., Chib, S., and Narasimhan, C. Markov chain monte carlo and models of consideration set and parameter heterogeneity. *Journal of Econometrics*, 89 (1):223–248, 1998.

Cramer, J. S. and Ridder, G. Pooling states in the multinomial logit model. *Journal of Econometrics*, 47(2-3):267–272, 1991.

Hauser, J. R. Consideration-set heuristics. *Journal of Business Research*, 67(8): 1688–1699, 2014.

Ho, T.-H. and Chong, J.-K. A parsimonious model of stockkeeping-unit choice. *Journal of Marketing Research*, 40(3):351–365, 2003.

Imai, K. and van Dyk, D. A. A bayesian analysis of the multinomial probit model using marginal data augmentation. *Journal of econometrics*, 124(2):311–334, 2005.

Ishwaran, H. and James, L. F. Approximate dirichlet process computing in finite normal mixtures. *Journal of computational and graphical statistics*, 2002.

Ishwaran, H. and Zarepour, M. Markov chain monte carlo in approximate dirichlet and beta two-parameter process hierarchical models. *Biometrika*, 87(2):371–390, 2000.

Jacobs, B. J., Donkers, B., and Fok, D. Model-based purchase predictions for large assortments. *Marketing Science*, 35(3):389–404, 2016.

Keane, M. P. A note on identification in the multinomial probit model. *Journal of Business & Economic Statistics*, 10(2):193–200, 1992.

Liu, Q. and Arora, N. Efficient choice designs for a consider-then-choose model. *Marketing Science*, 30(2):321–338, 2011.

Manzini, P. and Mariotti, M. Stochastic choice and consideration sets. *Econometrica*, 82(3):1153–1176, 2014.

McCulloch, R. and Rossi, P. E. An exact likelihood analysis of the multinomial probit model. *Journal of Econometrics*, 64(1):207–240, 1994.

McCulloch, R. E., Polson, N. G., and Rossi, P. E. A bayesian analysis of the multinomial probit model with fully identified parameters. *Journal of Econometrics*, 99(1):173–193, 2000.

McFadden, D. Modeling the choice of residential location. *Transportation Research Record*, (673), 1978.

Mehta, N., Rajiv, S., and Srinivasan, K. Price uncertainty and consumer search: A structural model of consideration set formation. *Marketing science*, 22(1): 58–84, 2003.

Naik, P., Wedel, M., Bacon, L., Bodapati, A., Bradlow, E., Kamakura, W., Kreulen, J., Lenk, P., Madigan, D. M., and Montgomery, A. Challenges and opportunities in high-dimensional choice data analyses. *Marketing Letters*, 19 (3-4):201–213, 2008.

Rossi, P. E., Allenby, G. M., and Robert, M. *Bayesian Statistics and Marketing*. John Wiley & Sons, Ltd, 2005.

Sethuraman, J. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

Terui, N., Ban, M., and Allenby, G. M. The effect of media advertising on brand consideration and choice. *Marketing Science*, 30(1):74–91, 2011.

Van Nierop, E., Bronnenberg, B., Paap, R., Wedel, M., and Franses, P. H. Retrieving unobserved consideration sets from household panel data. *Journal of Marketing Research*, 47(1):63–74, 2010.

Wasi, N., Keane, M. P., et al. Estimation of discrete choice models with many alternatives using random subsets of the full choice set: With an application to demand for frozen pizza. Technical report, Economics Group, Nuffield College, University of Oxford, 2012.

Zanutto, E. L. and Bradlow, E. T. Data pruning in consumer choice models. *Quantitative Marketing and Economics*, 4(3):267–287, 2006.

Appendix A Descriptive statistics explanatory variables empirical application

Table A1: Gross annual income of household categories

< 4.600	14.300 - 15.400	38.800 - 51.300	181.300 - 206.400
4.600 - 6.300	15.400 - 17.100	51.300 - 65.000	206.400 - 232.600
6.300 - 8.000	17.100 - 20.000	65.000 - 77.500	232.600 - 258.900
8.000 - 9.100	20.000 - 23.400	77.500 - 103.800	258.900 - 284.500
9.100 - 10.800	23.400 - 26.200	103.800 - 129.400	284.500 - 310.700
10.800 - 12.500	26.200 - 32.500	129.400 - 155.100	310.700 <
12.500 - 14.300	32.500 - 38.800	155.100 - 181.300	no response

This table shows the 28 categories of gross annual income of a household. The last category, no response, includes the households which do not know or do not want to say what their income is. The income categories are included as the log mean of each income group in the choice model. We correct for the no responses by including a dummy in the model for these households.

Table A2: Descriptive statistics explanatory variables

variable	mean	stdev	variable	1	0
Age	43.603	21.631	Income	774	2753
Income	8.473	4.525	Single	377	3150
			Children	1188	2339

This table shows the descriptive statistics of the explanatory variables. The first three columns show the mean and standard deviation of the continuous variables age and income of the respondent. The last three columns present frequency code for the included dummies in the model, no response for income of the household, whether it is a single household, or whether there are children in the household.

Appendix B Outcome categories and posterior results empirical application

Table B1: Outcome categories, frequency counts and posterior means PART 1

destinations	freq.	constant	age	inc	no resp	single	kids
France	576	0.000	0.000	0.000	0.000	0.000	0.000
Iceland	13	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Norway	39	0.458	0.003	-0.105	-1.266	0.029	-0.113
Sweden	23	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Finland	3	-0.843	-0.009	-0.023	-0.391	0.575	-0.661
Denmark	24	0.000	0.000	0.000	0.000	0.000	0.000
Ireland	16	-0.947	-0.002	-0.002	0.107	0.319	-0.541
United Kingdom	125	-0.698	-0.005	0.068	0.637	0.288	-0.552
Belgium	141	0.930	-0.004	-0.105	-1.177	0.173	-0.088
Luxembourg	27	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Germany	394	0.483	0.004	-0.034	-0.391	-0.033	-0.156
Switzerland	30	0.930	-0.004	-0.105	-1.177	0.173	-0.088
Austria	227	0.483	0.004	-0.034	-0.391	-0.033	-0.156
Poland	19	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Czech Republic	30	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Slovakia	14	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Hungary	24	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Romania	8	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Bulgaria	10	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Russia	4	-0.847	-0.009	-0.022	-0.382	0.579	-0.662
oc Eastern Europe	14	-0.611	-0.006	-0.037	-0.406	0.234	-0.645
Portugal	124	-0.698	-0.005	0.068	0.637	0.288	-0.552
Spain	439	0.483	0.004	-0.034	-0.391	-0.033	-0.156
Andorra	1	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Italy	254	0.483	0.004	-0.034	-0.391	-0.033	-0.156
Malta	13	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Slovenia	1	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Croatia	62	0.282	0.005	-0.093	-1.264	-0.065	-0.173
Greece	158	0.483	0.004	-0.034	-0.391	-0.033	-0.156
oc Southern Europe	33	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Morocco	12	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Tunesia	4	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Egypt	30	0.312	-0.001	-0.096	-1.125	0.002	-0.289

This table shows the first 33 outcome categories and the frequency counts for each category in the first two columns. Other countries in is abbreviated to oc. The last six columns show for each outcome category the posterior means for the constant, and the parameters for age, income, no response to income, single household, and household with children.

Table B2: Outcome categories, frequency counts and posterior means PART 2

destinations	freq.	constant	age	inc	no resp	single	kids
Kenya	6	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc Eastern Africa	12	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Gambia	4	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc West Africa	2	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
South Africa	14	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc South Africa	6	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Cyprus	9	-0.947	-0.002	-0.002	0.107	0.319	-0.541
Israel	1	-0.847	-0.009	-0.022	-0.382	0.579	-0.662
Turkey	147	0.000	0.000	0.000	0.000	0.000	0.000
Jordan	3	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc West Asia	2	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
India	0	-0.769	-0.006	-0.032	-0.489	0.378	-0.752
Sri Lanka	8	-0.498	-0.009	-0.012	0.031	0.011	-0.982
oc South Asia	2	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
China	8	-0.947	-0.002	-0.002	0.107	0.319	-0.541
oc East Asia	13	-0.947	-0.002	-0.002	0.107	0.319	-0.541
Indonesia	60	-0.773	-0.007	0.059	0.529	0.033	-0.770
Thailand	13	0.312	-0.001	-0.096	-1.125	0.002	-0.289
Malaysia	5	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc South-east Asia	11	-0.460	-0.010	-0.013	0.013	-0.002	-0.999
oc Asia	4	-0.705	-0.006	-0.033	-0.426	0.293	-0.732
Australia	30	-0.460	-0.010	-0.013	0.013	-0.002	-0.999
New Zealand	4	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
Canada	51	0.000	0.000	0.000	0.000	0.000	0.000
United States	111	-0.698	-0.005	0.068	0.637	0.288	-0.552
Netherlands Antilles	44	-0.773	-0.007	0.059	0.529	0.033	-0.770
Dominican Republic	2	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc Carribean	9	-0.947	-0.002	-0.002	0.108	0.319	-0.541
Mexico	11	-0.612	-0.006	-0.037	-0.408	0.212	-0.657
oc Central America	9	-0.665	-0.009	-0.024	-0.243	0.402	-0.599
Brazil	14	-0.837	0.008	-0.104	-0.755	-0.805	-0.394
oc South America	20	-0.523	-0.009	-0.010	0.052	0.013	-0.969
oc world	0	-0.612	-0.006	-0.037	-0.408	0.212	-0.657

This table shows the last 33 outcome categories and the frequency counts for each category in the first two columns. Other countries in is abbreviated to oc. The last six columns show for each outcome category the posterior means for the constant, and the parameters for age, income, no response to income, single household, and household with children.