MONASH University

MONASH BUSINESS SCHOOL

ISSN 1440-771X

Department of Econometrics and Business Statistics

http://business.monash.edu/econometrics-and-business-statistics/research/publications

A high-dimensional multinomial choice model

Didier Nibbering

September 2019

Working Paper 19/19

A high-dimensional multinomial choice model

Didier Nibbering[*]

Department of Econometrics and Business Statistics, Monash University

February 25, 2019

Abstract

The number of parameters in a standard multinomial choice model increases linearly with the number of choice alternatives and number of explanatory variables. Since many modern applications involve large choice sets with categorical explanatory variables, which enter the model as large sets of binary dummies, the number of parameters easily approaches the sample size. This paper proposes a new method for data-driven parameter clustering over outcome categories and explanatory dummy categories in a multinomial probit setting. A Dirichlet process mixture encourages parameters to cluster over the categories, which favours a parsimonious model specification without a priori imposing model restrictions. An application to a dataset of holiday destinations shows a decrease in parameter uncertainty, an enhancement of the parameter interpretability, and an increase in predictive performance, relative to a standard multinomial choice model.

Keywords: large choice sets, Dirichlet process prior, multinomial probit model, high-dimensional models

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

[*] e-mail: didier.nibbering@monash.edu. I would like to thank Richard Paap, Michel van der Wel, Dennis Fok, Tom Boot, conference participants at the 4th Annual Conference of the International Association for Applied Econometrics (2017), 8th European Seminar on Bayesian Econometrics (2017), European Winter Meeting of the Econometric Society (2017), NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics (2018), and seminar participants at the Tinbergen Institute for helpful discussions. I thank NBTC Holland Marketing for access to the survey data from the 'ContinuVakantieOnderzoek'.

1 Introduction

Many multinomial choice problems involve large choice sets relative to the number of observations. Video streaming providers have choice data on millions of movies, the assortment size in grocery stores exceeds thousands of products, and we have the choice between almost every place in the world as holiday destination.

To identify factors that explain observed choices, choice behavior is related to characteristics of decision makers. Streaming providers take past ratings for different kind of genres of their subscribers into account before recommending movies. With the advent of loyalty cards, grocery stores know exactly in which neighborhoods their costumers live. Online travel agencies ask for your household composition before offering travel deals. These individual characteristics divide the sample of decision makers into a large number of different categories.

Multinomial choice models help to understand the relation between discrete choices and the characteristics of the decision makers. Since the parameters in these discrete choice models are alternative-specific, the number of parameters increases linearly with the size of the choice set. Furthermore, when the explanatory variables describe categorical characteristics, these variables enter the model as sets of dummies, with for each category a dummy variable. With several categorical variables and large numbers of categories, the number of parameters needed to specify the effect on one choice alternative is already large. When both the number of choice alternatives and the number of explanatory categories is large, the number of parameters easily approaches the number of observations.

This paper proposes a Bayesian method to manage the number of parameters in high-dimensional multinomial choice models in a data-driven way. A two-way Dirichlet process prior on the model parameters of a multinomial probit model encourages the alternative-specific parameters to cluster over both outcome and explanatory categories. With positive probability, the two-way mixture choice model reduces the high-dimensional parameter space in both directions. The result is a decrease in parameter uncertainty and an enhancement of the parameter interpretability, without imposing any model restrictions.

Although pooling of categories is ubiquitous in practice, we are, to the best of our knowledge, the first to estimate from the data which categories can be pooled together, for both dependent and independent categorical variables. We set up a Gibbs sampler that draws model parameters clustered over outcome and explanatory categories, and draws two-way cluster assignments over both dimensions. Since we can formulate the multinomial probit model in terms of latent normally distributed utilities (Albert and Chib, 1993), the cluster assignments can be sampled according to the sample steps developed for mixtures of normals of Ishwaran and James (2002). By jointly estimating the number of clusters, cluster assignments, and model parameters, the posterior parameter distributions incorporate the parameter uncertainty together with the uncertainty in the number of clusters, which is ignored by fixing a priori the number of clusters. The estimated model parameters retain their interpretation as in a standard multinomial choice model, and prior distributions can be parametrized according to prior beliefs about the number of distinct effects over outcome and explanatory categories.

In an empirical application we estimate the effect of household composition on holiday destinations. We apply the two-way mixture choice model to survey data on holiday behavior of Dutch households, and estimate the effect of 21 explanatory variables describing household characteristics on the choice out of 49 holiday destinations. The estimated holiday destination clustering is very different from an ad hoc grouping based on, for instance, geographical location. We show how the estimated parameter clustering over households and holiday destinations reduces the uncertainty about household preferences, provides insights in the relation between household characteristics and holiday preferences, and find a substantial increase in predictive performance relative to a standard choice model.

When confronted by a large number of alternatives, researchers commonly focus on a subset of alternatives, or alternatives are a priori aggregated to a higher level (Zanutto and Bradlow, 2006; Carson and Louviere, 2014). This is not a solution when all available categories are of interest. Cramer and Ridder (1991) propose a statistical test for pooling outcome categories. However, testing for all different combinations of subsets is computationally expensive and the order of tests can change the final clustering. At the cost of departing from the standard discrete choice model parameter interpretation, Ho and Chong (2003) and Jacobs et al. (2016) circumvent the pooling problem by introducing an additional set of latent variables. Instead of estimating separate parameters for each choice alternative, the explanatory variables influence the choice probabilities via a small set of latent variables. Chiong and Shum (2018) analyze large choice sets with aggregated choice data, ruling out estimates or predictions on the decision maker level.

Large sets of explanatory categories are, similar to choice alternatives, often clustered on expert opinion to ease the curse of dimensionality. Evidently, this leads to suboptimal results when the expert is wrong. More recently, regularization techniques for high-dimensional regressor matrices, such as the lasso introduced by Tibshirani (1996), are also applied to categorical data. In addition to shrinkage or selection, it is for a categorical explanatory variable also of interest which categories should be distinguished when modelling the effect on the outcome variable (Tutz and Gertheiss, 2016). Bondell and Reich (2009) and Gertheiss et al. (2010) show that by choosing a specific functional form for the penalty in the lasso, categories are clustered to a smaller set of dummies. Although these methods are tailored to the categorical nature of the data, the relation between the lasso penalty parameter and the number of distinguished categories is opaque.

The potential of the Dirichlet process prior has gained an increasing amount of attention. The most popular application of the prior is the modelling of unknown error distributions, without resorting to strong parametric assumptions. Hirano (2002) puts a Dirichlet process prior on the error distribution in dynamic panel data models, Van Hasselt (2011) in sample selection models, and Chib and Hamilton (2002), Conley et al. (2008) and Wiesenfarth et al. (2014) in instrumental variable models. On the other hand, Dirichlet process priors are used to model parameter heterogeneity. Hu et al. (2015) specify the prior on the model parameters in an instrumental variable model to allow for heterogeneity in treatment effects. Bauwens et al. (2017) use the prior to model time-variation in the parameters of autoregressive moving average models. Burda et al. (2008) models unobserved heterogeneity across individuals by a Dirichlet process prior on individual-specific parameters in a choice model. Bernstein et al. (2018) use the prior for the dynamic estimation of clusters of customers with similar preferences. We employ the properties of a Dirichlet process prior to embattle choice models for high-dimensional choice sets. Instead of mixing over individuals or over time, the prior clusters parameters over choice alternatives and explanatory categories.

The outline of the remainder of this paper is as follows. Section 2 discusses the model specification and Section 3 explains the Bayesian inference methods. Section 4 applies the two-way mixture model to survey data on holiday destinations. We conclude with a discussion in Section 5. Appendix A discusses details on the inference procedure, Appendix B provides insights in model performance with a numerical experiment, and Appendix C gives an overview of the data used in the empirical application.

2 Model specification

This section discusses the specification of a high-dimensional multinomial choice model. Section 2.1 introduces the baseline specification of a multinomial probit model. Section 2.2 shows how parameters are clustered over the categories in the categorical dependent and independent variables in this model. Section 2.3 discusses the technique that drives the clustering, the Dirichlet process prior.

2.1 Multinomial probit model

Let y_i be an observable unordered random categorical variable, such that $y_i \in \{1, 2, \ldots, J\}$, with J the number of choice alternatives, and $i = 1, \ldots, N$, with N the number of individuals. Let x_i be a K-dimensional vector with explanatory variables, potentially with dummy coded categorical variables. As is common for multinomial choice models, we introduce latent utilities driving the decisions. Let $z_i = (z_{i1}, \ldots, z_{iJ})'$ be a $J \times 1$ vector of continuous latent random 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 utilities are modeled as

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

where $\alpha = (\alpha_1, \ldots, \alpha_J)'$ is a J-dimensional vector of intercepts, $\beta = (\beta_1, \ldots, \beta_J)'$ is a $J \times K$ matrix of coefficients, and ε_i an independent normally distributed disturbance vector with covariance matrix Σ. We now have defined the conditional density $f(y_i|x_i, \alpha, \beta, \Sigma)$, where the covariates in x_i are constant across different outcome categories, but the K-dimensional model parameter vectors β_j, $j = 1, \ldots, J$, vary over the outcome categories.

The parameters α_j, β_j and Σ in the multinomial probit model specified in (1) and (2) are not identified (Bunch, 1991). There are two parameter identification problems. First, $y_i(z_i + c) = y_i(z_i)$ for each scalar c. To overcome this additive redundancy we set $\alpha_1 = 0$ and $\beta_1 = 0$. Second, we still have $y_i(cz_i) = y_i(z_i)$ for each positive scalar c, even if the aforementioned restriction is imposed. We follow Gilbride and Allenby (2004) and Terui et al. (2011) and set the covariance matrix Σ in (2) to be the identity matrix. This restriction identifies the model parameters and avoids covariance parameter estimation problems when J is large.

A conventional multiplicative identifying assumption is to only 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) report the posterior of the model parameters up to a scaling factor. 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, in the absence of alternative-specific explanatory variables, parameter identification in 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, especially in the high-dimensional settings we study in this paper.

2.2 Parameter clustering over categories

When the choice set is large, the number of parameters in the $J \times K$ matrix β easily approaches the number of observations. Large numbers of parameters amplify overfitting concerns, increase parameter uncertainty, and make it a difficult exercise to extract useful insights. For the data to be informative on the parameters without additional restrictions, the number of outcome categories and the number of explanatory variables need to be relatively small.

Two features of many large scale empirical applications of choice models exacerbate the curse of dimensionality. First, the observed choices $y_1, \ldots, y_N$ often are not evenly distributed over the choice set. This results in a small number of observed choices to estimate the parameters β_j for the least chosen alternatives j, even for large N relative to J. Second, the individual choice behavior is usually explained by, among other variables, categorical variables indicating characteristics of individuals. These categorical variables are implemented by means of dummies, resulting in sets of binary variables for each explanatory category. Therefore, the number of explanatory variables K can become large in models with categorical variables consisting of many explanatory categories.

When subsets of categories can be treated as a single category, this parsimonious model is preferred. Section 2.2.1 discusses parameter clustering over outcome categories, Section 2.2.2 over explanatory categories, and Section 2.2.3 over both dimensions.

2.2.1 Parameter clustering over outcome categories

The latent utility model in (2) can be written as

$$z_{ij} = \alpha_j + \beta_j' x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{3}$$

where the vector β_j contains alternative-specific parameters. Equivalently, we can say that the parameters in β_j vary over an infinite number of clusters, where the number of clusters equals the total number of choice alternatives J when each choice alternative has a different parameter vector. Within the clusters the parameters are assumed to be identical, but across clusters the parameters are allowed to be different.

The cluster representation of the latent utility model in (3) is

$$z_{ij} = \alpha_j + \tilde{\beta}_{C_j}' x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{4}$$

where $\beta_j = \tilde{\beta}_{C_j}$ can vary over $L_J \to \infty$ clusters. The classification variables $C_j \in \{1, \dots, L_J\}$ take integer values indicating the cluster for choice category j, and identify the corresponding cluster parameter vector $\tilde{\beta}_{C_j}$.

The model in (4) imposes the parameter clustering over choice alternatives to be the same for each individual. Although this seems restrictive at first sight, the clustering does not imposes the same expected utility ordering for each individual. Since the utilities are conditional on individual-specific characteristics in x_i, the expected preference ordering over outcome categories is also individual-specific. However, when repeated observations per individual are available, we can easily extend (4) to a more flexible model with individual-specific clustering.

2.2.2 Parameter clustering over explanatory categories

To cluster over categories within a categorical explanatory variable, we make an explicit distinction in the regressor vector $x_i = (w_i', d_i')'$. The K_d dummies in d_i correspond to the first K_d of the $K_d + 1$ categories in the categorical explanatory variable. The intercepts in α capture the effect of the last explanatory category. The vector w_i contains the K_w remaining explanatory variables. We rewrite the model in (3) to

$$z_{ij} = \alpha_j + \beta_j' x_i + \varepsilon_{ij} = \alpha_j + \gamma_j' w_i + \kappa_j' d_i + \varepsilon_{ij}, \tag{5}$$

where $\beta_j = (\gamma_j', \kappa_j')'$ and where the parameter values in $\kappa_j = (\kappa_{j,1}, \dots, \kappa_{j,K_d})$ correspond to the dummy categories in $d_i = (d_{i,1}, \dots, d_{i,K_d})$. We cluster the dummy parameters over the categories of only one categorical explanatory variable in (5). However, the methods can easily be extended to account for parameter clustering over multiple explanatory categorical variables.

We let the explanatory dummy parameters vary over an infinite number of clusters. The formulation of the latent utility model in (5) conditional on the classification variables is

$$z_{ij} = \alpha_j + \gamma_j' w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \dots, J, \tag{6}$$

where $\kappa_{jk} = \tilde{\kappa}_{j,D_k}$ can vary over $L_D \to \infty$ clusters. The classification variables $D_k \in \{1, \ldots, L_D\}$ take integer values indicating the cluster for explanatory category k. Within a cluster l, dummies have identical parameter values $\tilde{\kappa}_{jl}$ and are equivalently aggregated to a new dummy variable. As a result, the explanatory categories within one cluster have the same effect on the dependent variable and we have a smaller set of dummies.

The dummy parameter clustering in (6) perfectly fits categorical variables without a natural ordering, such as profession. However, the modelling framework does not take ordering in the explanatory categories into account. Ordered explanatory categories, for instance income categories, fit in as well but could be handled more efficiently when the ranking in the categories can be taken into account.

2.2.3 Two-way parameter clustering

Combining parameter clustering over outcome categories in (4) with parameter clustering over explanatory categories in (6) results in two-way parameter clustering,

$$z_{ij} = \alpha_j + \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1, \ldots, J, \tag{7}$$

where $\gamma_j = \tilde{\gamma}_{C_j}$ can vary over L_J clusters and $\kappa_{jk} = \tilde{\kappa}_{C_j,D_k}$ over $L_J \times L_D$ clusters.

The parameter clustering in (7) over the outcome category dimension is unconditional on the clustering over the explanatory category dimension. This means that the clustering over dummy parameters is the same for each outcome category cluster. Allowing for conditional clustering, where each cluster of outcome categories may have another division of explanatory categories, results in an overly flexible model specification which causes difficulties in parameter estimation and parameter interpretation.

2.3 Dirichlet process mixture model

The key to our parameter clustering approach is the specification of a cluster assignment probability distribution for each category. The probability distribution is modelled by a Dirichlet process mixture model that implicitly integrates out the cluster probabilities, while allowing for as many clusters as categories. Since there is a positive probability that two categories share a cluster, the Dirichlet process mixture encourages a parsimonious model without imposing any model restrictions.

2.3.1 Dirichlet process prior

A data-driven parameter clustering approach is obtained by specifying a Dirichlet process prior for the parameter vector β_j in (3),

$$\begin{aligned} \beta_j|P &\sim P, \\ P|\lambda_J, H &\sim DP(\lambda_J, H_\beta), \end{aligned} \tag{8}$$

where the prior of β_j is a random distribution P generated by a Dirichlet process. Conditionally on P, the parameter vectors β_j, $j = 1, \dots, J$, are independently and identically distributed. The Dirichlet process $DP(\lambda_J, H_\beta)$, has a positive scalar concentration parameter λ_J and continuous base distribution H_β.

The expectation over the Dirichlet process equals the base distribution, and the concentration parameter governs the dispersion around the base distribution. When λ_J is large, the distributions P and H_β are more similar. Since P is a discrete random distribution, there is a positive probability that different β_j's take the exact same value. A cluster of outcome categories is defined as the choice alternatives with identical parameter vectors β_j. Therefore, the model in (8) is known as a Dirichlet process mixture model, which in this case clusters over choice alternatives.

A standard multinomial choice model puts a prior on the parameters that assumes that each β_j is independent and identically distributed; $\beta_j \sim iid\, H_\beta$ for $j = 1, \dots, J$. In this case, the parameters values β_j are unique. A Dirichlet process prior also allows the parameters to vary over j. However, the prior clusters similar categories into groups with unique values of β_j with a positive probability.

2.3.2 Stick-breaking representation

Sethuraman (1994) shows that a Dirichlet process prior is equivalently formulated by the stick-breaking representation,

$$P = \sum_{l=1}^{L_J} p_l \delta(\tilde{\beta}_l), \qquad \tilde{\beta}_l \sim H_\beta, \tag{9}$$

where $L_J \to \infty$ and $\delta(\tilde{\beta}_l)$ denotes a unit-mass measure concentrated at $\tilde{\beta}_l$. The Dirichlet process is a distribution over independent and identically distributed draws from the base distribution, with random weights

$$p_1 = V_1, \quad p_l = (1 - V_1)(1 - V_2) \dots (1 - V_{l-1})V_l, \quad l = 2, \dots, L_J, \tag{10}$$

where $V_l \sim \text{Beta}(1, \lambda_J)$. This process is also written as $p = \{p_l\}_{l=1}^{L_J} \sim \text{stick}(\lambda_J)$. Since $\sum_{l=1}^{L_J} p_l = 1$, it follows that p can be interpreted as probabilities, and P is a distribution over discrete probability measures.

The construction of the weights p_l in (10) is named after the process of iteratively breaking up a stick into pieces. Starting with a unit-length stick, in each step we break off a random proportion of the remaining stick. When we write (10) as

$$p_l = V_l \prod_{k=1}^{l-1} (1 - V_k), \tag{11}$$

we can interpret V_l as the proportion of the remaining stick which has length p_l. After breaking off the first $l - 1$ pieces, the length of the remainder of the stick is $\prod_{k=1}^{l-1}(1 - V_k)$. Since $\text{E}[V_l] = \frac{1}{1+\lambda_J}$, a small λ_J results on average in a few large sticks, and the lengths of the remaining sticks are close to zero. For a large value for λ_J, the weights in p are more evenly distributed.

2.3.3 Mixture model over outcome categories

The Dirichlet process mixture model in (8) can be equivalently formulated by means of the classification variables $C = (C_1, \dots, C_J)$ in (4). Using the stick-breaking representation of the Dirichlet process prior in (9), we have

$$\begin{gathered} z_{ij} = \alpha_j + \tilde{\beta}'_{C_j} x_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \\ C_j | p \sim \sum_{l=1}^{L_J} p_l \delta(l), \quad p \sim \text{stick}(\lambda_J), \quad \tilde{\beta}_l \sim H_\beta. \end{gathered} \tag{12}$$

The probability that an outcome category is assigned to cluster l is denoted by p_l. The conditional distribution of z_{ij} now takes the form of a mixture distribution over the outcome categories with random weights $p = (p_1, \ldots, p_{L_J})$,

$$f(z_{ij}|x_i, \alpha_j, \tilde{\beta}_1, \ldots, \tilde{\beta}_{L_J}, p) = \sum_{l=1}^{L_J} p_l f_N(z_{ij}|\alpha_j + \tilde{\beta}'_l x_i, 1), \tag{13}$$

where $f_N(x|\mu, \sigma^2)$ is a normal density with expectation μ and variance σ^2 evaluated at x. The conditional distribution in (13) is an infinite mixture of normal distributions.

The Dirichlet process mixture model over choice alternatives infers whether the parameters of a subset of categories can be treated as a single parameter, or whether the alternative-specific parameters are significantly different to distribute them over different clusters. Even when there are differences between categories, but there is not enough power to distinguish all differences between the values in the category-specific parameter vectors, it may be that the efficiency gain of clustering still outweighs the loss in accuracy. Therefore, the Dirichlet process mixture model only introduces a new parameter vector for a outcome category when this category is significantly different from the other ones.

2.3.4 Mixture model over explanatory categories

Along the same lines as for outcome categories, we specify a Dirichlet process prior for the explanatory dummy parameters in (5),

$$\begin{gathered} z_i = \alpha + \gamma w_i + \sum_{k=1}^{K_d} \kappa_{1:J,k} d_{ik} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_J), \\ \gamma \sim H_\gamma, \quad \kappa_{1:J,k}|Q \sim Q, \quad Q|\lambda_D, H_\kappa \sim DP(\lambda_D, H_\kappa), \end{gathered} \tag{14}$$

where $\gamma = (\gamma_1, \dots, \gamma_J)'$ and $\kappa_{1:J,k} = (\kappa_{1k}, \dots, \kappa_{Jk})'$, H_κ is the base distribution of the parameters $\kappa_{1:J,k}$, and H_γ the prior distribution for γ. In the same way as for the outcome category cluster probabilities, we let the explanatory category cluster probabilities $q = \{q_l\}_{l=1}^{L_D} \sim \text{stick}(\lambda_D)$. The stick breaking representation conditional on the classification vector $D = (D_1, \dots, D_{K_d})$ is

$$\begin{aligned} z_i &= \alpha + \gamma w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{1:J,D_k} d_{ik} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, I_J), \\ D_k|q &\sim \sum_{l=1}^{L_D} q_l \delta(l), \quad q \sim \text{stick}(\lambda_D), \quad \gamma \sim H_\gamma, \quad \tilde{\kappa}_{1:J,l} \sim H_\kappa, \end{aligned} \tag{15}$$

where $\tilde{\kappa}_{1:J,l} = (\tilde{\kappa}_{1l}, \dots, \tilde{\kappa}_{Jl})'$, and q_l is the probability that an explanatory category is assigned to cluster l. The elements of the explanatory cluster assignment probability vector $q = (q_1, \dots, q_{L_D})$ add up to one, $\sum_{l=1}^{L_D} q_l = 1$.

From (15) follows the specification of the Dirichlet process mixture model that mixes over the parameter values corresponding to the dummies variables d_i,

$$f(z_i|x_i, \alpha, \gamma, \tilde{\kappa}_{1:J,1}, \dots, \tilde{\kappa}_{1:J,L_D}, q) = \sum_{l=1}^{L_D} q_l f_N(z_{ij}|\alpha + \gamma w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{1:J,l} d_{ik}, 1). \tag{16}$$

The Dirichlet process mixture model over explanatory categories in (16) clusters dummy parameters together. A subset of explanatory dummy categories with identical parameters is equivalent to aggregating the corresponding explanatory dummy variables.

2.3.5 Two-way mixture model

We specify a mixture model for the two-way parameter clustering in Section 2.2.3 by combining the mixture model over outcome categories in (12) with a mixture model over explanatory categories in (15). The result is a two-way Dirichlet process mixture model

$$\begin{gathered}
z_{ij} = \alpha_j + \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1), \quad j = 1,\dots,J, \\
C_j|p \sim \sum_{l=1}^{L_J} p_l \delta(l), \quad p \sim \text{stick}(\lambda_J), \quad D_k|q \sim \sum_{k=1}^{L_D} q_k \delta(k), \quad q \sim \text{stick}(\lambda_D), \\
\tilde{\gamma}_l \sim H_\gamma, \quad \tilde{\kappa}_{lk} \sim H_{\kappa_j}, \quad l = 1,\dots,L_J, \quad k = 1,\dots,L_D.
\end{gathered} \tag{17}$$

The Dirichlet process mixture models in (12), (15), and (17), are tailored for a high-dimensional multinomial probit model. However, they can easily be extended to a mix of a multinomial and conditional choice model by adding a conditional part to (2), in which the covariates vary across different outcome categories, but the model parameters are constant over outcome categories. The methods are of less interest to the conditional choice model itself, since the parameters do not grow in the number of choice alternatives. The same holds for ordered choice models, in which only the intercepts are alternative-specific. Another interesting application of the cluster methods is the rank ordered model, in which the model parameters are also alternative-specific.

3 Bayesian inference

To estimate the posterior distributions of the parameters in β, we approximate the two-way Dirichlet process mixture model by truncating the Dirichlet processes at the Lth term by setting $V_L = 1$ for a finite number L. The Gibbs sampler for the truncated Dirichlet process is simpler than corresponding samplers for the full Dirichlet process, while displaying favorable mixing properties (Ishwaran and James, 2002). Section 3.1 discusses the prior distributions and Section 3.2 the posterior distribution. Appendix A.1 shows that the approximation error resulting from the truncation is small and Appendix A.2 shows how the prior distributions can be parametrized according to prior beliefs about the number of distinct effects over outcome and explanatory categories. Appendix A.3 provides a detailed outline of the posterior sampling algorithm and Appendix A.4 of the predictive sampling algorithm.

3.1 Prior distributions

The Dirichlet process mixture model is defined by a Dirichlet process prior on the parameters β. To complete the prior specification for β, we specify the base distribution H_β,

$$\beta_{1k} \sim \mathcal{N}(0,0) \text{ and } \beta_{jk}|\sigma_\beta^2 \sim \mathcal{N}(0,\sigma_\beta^2), \tag{18}$$

where $j = 2,\ldots,J$, $k = 1,\ldots,K$, and $\sigma_\beta \in \mathcal{R}^+$. Note that the normal distribution turns into the Dirac delta function $\delta(0)$ when the variance is zero. We let the data determine the number of clusters by treating the concentration parameters as unknown with a prior distribution,

$$\lambda_J|\eta_{J1},\eta_{J2} \sim \text{Gamma}(\eta_{J1},\eta_{J2}), \quad \lambda_D|\eta_{D1},\eta_{D2} \sim \text{Gamma}(\eta_{D1},\eta_{D2}), \tag{19}$$

where $\text{Gamma}(\eta_{\cdot 1},\eta_{\cdot 2})$ denotes a gamma distribution with mean $\eta_{\cdot 1}/\eta_{\cdot 2}$. The values $(\eta_{\cdot 1},\eta_{\cdot 2}) \in \mathcal{R}^+$ directly effect the number of estimated clusters through the concentration parameter, where larger values for $\lambda_\cdot$ encourage more distinct values for the coefficients. We set $\eta_{\cdot 1}/\eta_{\cdot 2}$ equal to the value of the concentration parameter that matches the prior belief on the number of clusters as discussed in Appendix A.2, and use $\eta_{\cdot 2}$ to govern the dispersion around the mean.

We conclude with the prior specification for the intercept parameters α,

$$\alpha_1 \sim \mathcal{N}(0,0) \text{ and } \alpha_j|\sigma_\alpha^2 \sim \mathcal{N}(0,\sigma_\alpha^2), \tag{20}$$

where $j = 2,\ldots,J$, and $\sigma_\alpha \in \mathcal{R}^+$.

3.2 Posterior distribution

To estimate the posterior distributions of the parameters, we rely on a Markov Chain Monte Carlo sampler with data augmentation. The representations of the mixture models in (12) and (15) condition on the choice alternative classification variable C and the explanatory dummy category classification variable D, respectively. Using these representations of the model allows for clustering over both the outcome category and the explanatory category dimension, by simulating the latent classification variables alongside the model parameters in α and β, and the cluster probabilities p and q.

In each iteration of the Gibbs sampler, we sample the parameters α, β, p, and q together with the latent classification variables C and D from their full conditional distributions, given the data $y = (y_1, \dots, y_N)'$ and $x = (x_1, \dots, x_N)'$. The Markov Chain Monte Carlo simulation scheme is as follows:

1. Sample $z|\alpha, \tilde{\beta}, C, D, y, x$
2. Sample $\alpha|\tilde{\beta}, C, \sigma_\alpha, z, x$
3. Sample $\tilde{\beta}|C, D, \sigma_\beta, \alpha, z, x$
4. Sample $C|p, \alpha, \tilde{\beta}, D, z, x$
5. Sample $D|q, \alpha, \tilde{\beta}, C, z, x$
6. Sample $p|C, \lambda_J$ and $q|D, \lambda_D$
7. Sample $\lambda_J|p, \eta_{J1}, \eta_{J2}$ and $\lambda_D|q, \eta_{D1}, \eta_{D2}$

The first sampling step distinguishes the sampling algorithm for the multinomial probit model from a normal mixture model. Since the multinomial probit model can be represented by a set of Gaussian latent variables, as we show in (2), sampling the latent variables $z = (z_1, \dots, z_N)'$ conditional on the observed choices in y is sufficient (Albert and Chib, 1993). Since z contains continuous normally distributed variables, it can serve as dependent variable in the sampling steps of Ishwaran and James (2002).

The intercept and coefficient parameters are sampled conditional on z. The parameters in the parameter matrix $\tilde{\beta}$, with rows $\tilde{\beta}_l = (\tilde{\gamma}_l', \tilde{\kappa}_{l,1}, \dots, \tilde{\kappa}_{l,L_D})$, are sampled in the third step, which extends the sampler of Ishwaran and James (2002) in two directions. First, their sample algorithm is developed for normal mixtures over the observations $i = 1, \dots, N$, which is relatively straightforward since clusters of observations are independent of each other. We sample parameter values for clusters over the dimensions $j = 1, \dots, J$ and $k = 1, \dots, K_d$. Second, we extend sampling model parameters over one-way clusters to two-way clustering, by sampling the model parameters simultaneously over the outcome and explanatory category clusters.

The fourth and fifth sampling steps draw the classification vectors. For identification purposes, the first outcome category is in every iteration of the sampler assigned to the first cluster, in which the parameter values $\tilde{\beta}_1$ are equal to zero. Since the categorical explanatory regressors d_i may be correlated with explanatory variables in w_i, there is potential dependence between parameters corresponding to different category dummies, which we take into account when sampling the classification variables D.

The probabilities of each cluster of outcome categories and the probabilities of each cluster of explanatory categories are sampled in the sixth step, and finally we resample the concentration parameters. Appendix A.3 provides a detailed outline of the sampling steps.

4 Empirical application

This section estimates the effect of household composition on holiday destinations using survey data from a Dutch market research company. The market research company is interested how preferences for holiday destinations differ across household types.

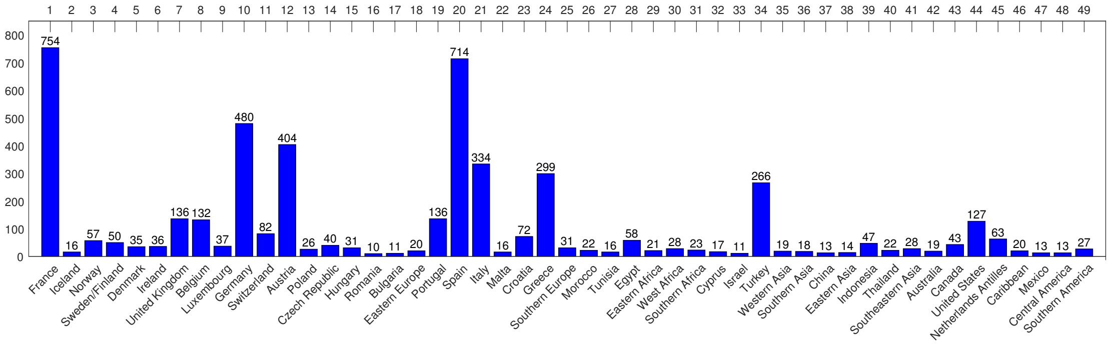
Due to the large number of choice categories and explanatory categories, this question is well-suited to be analyzed by the two-way mixture model. Dutch holidaymakers visit a large number of different holiday destinations each year. Households can be grouped in different categories with single households, couples with children of different age groups, or families of adults. Since the household composition is a categorical variable, it enters the analysis as a set of dummy variables. Estimating the effects of these household dummies on the holiday destination categories in a conventional choice model, results in a large number of parameter estimates. The large amount of parameter uncertainty and the large number of parameter estimates to be interpreted make it difficult to extract an answer to the research question.

4.1 Data

The data set consists of details of all reported holidays undertaken in 2015 by 6512 Dutch respondents and their individual characteristics. Among other things, respondents were asked to which country or region they have been for holidays and for how long. Since decision processes of households differ between short breaks and long vacations, we analyze the 4907 holidays with a foreign destination of more than seven days. Jointly analyzing the decision process for the 1881 domestic holidays and the 4907 foreign holidays asks for a baseline inflated choice model, which is outside the scope of this paper.

The respondents could select their foreign holiday destination from 77 categories in the survey, from which the market research company grouped countries of certain regions into one category. Subsequently, we grouped categories in the survey answers to end up with a minimum of ten observations per category. Categories which are never chosen by respondents are deleted. Appendix C.1 shows the countries per holiday destination choice category. We set the most frequent chosen holiday destination, which is France, as the base category. Figure 1 shows the frequency counts for the 49 categories in the resulting dependent variable. The overall pattern in the survey answers is representative for the Dutch holiday market.

Figure 1: Frequency counts choice categories

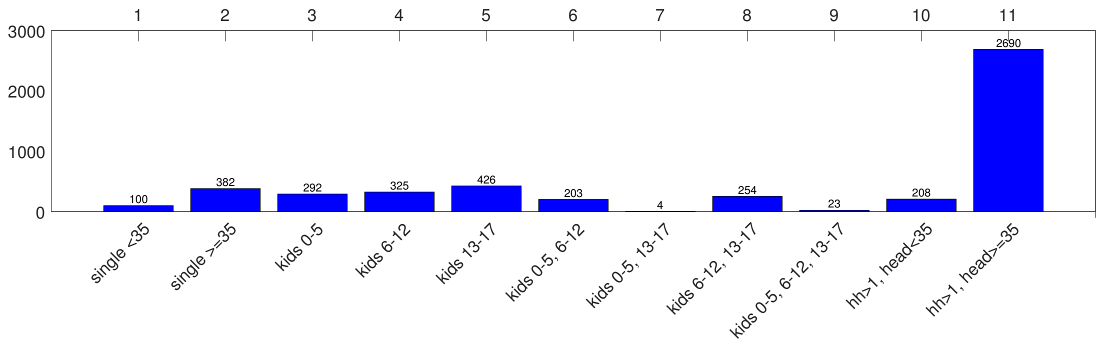


This figure shows the frequency counts for the categorical dependent variable. The categories represent destinations of foreign holidays of more than seven days of Dutch households.

The survey asked the respondents to select their household composition out of eleven categories. The first two categories distinguish singles under 35 from singles above 35. The third till ninth category describe households with children. Kids are divided among the age groups 0-5, 6-12, and 13-17, and four categories describe all possible combinations of these age groups in a family. The final two categories contain households of two or more persons in which everyone is 18 years or older, with the head of the household under 35 or older than 35. Figure 2 shows the frequency counts for the dummy categories.

In addition to the set of household composition dummy variables, we include ten control variables. We control for the income of the household, which is measured as a categorical variable, by the standardized logarithm of the maximum of the income category of the household in a continuous variable. A dummy variable corrects for respondents who do not want to say or do not know their income. The set of controls is completed by dummies indicating respondents who are retired, are student, own a moving holiday accommodation, own a fixed holiday accommodation, and are in a specific social class. Appendix C.2 explains the control variables in more detail and provides descriptive statistics.

Figure 2: Frequency counts household categories



This figure shows the frequency counts for the categorical explanatory variable. The categories represent the household compositions of the survey respondents for each holiday.

4.2 Modelling choices

We estimate the parameters β in the multinomial probit model defined in (1) and (2) on the first 4000 holidays and use the remaining 907 holidays for out-of-sample analysis. In the standard multinomial probit model, all parameters in the $J \times K$ matrix β are unique, which amounts to $49 \times 21 = 1029$ parameters with only 4000 observations on a nominal scale. To decrease parameter uncertainty and increase interpretability of the results, we cluster over both dimensions of the parameter matrix β in the two-way mixture model (17). The parameters are sampled as discussed in Section 3.

The truncation level of the number of potential choice category clusters is set equal to $L_J = 25$. Since the number of choice categories is $J = 49 > L_J$, we are charged with an approximation error. Following the procedure in Appendix A.1, we find that the expectation and the variance of the aggregated higher order probabilities equal 0.038 and 6.926×10^{-4} for the sampled α_J in the last iteration of the Gibbs sampler. We do not truncate the number of potential dummy category clusters, $L_D = K_d = 10$, which means that we specify a full Dirichlet process for the explanatory categories.

We follow Appendix A.2 in choosing the parameter values in the prior distributions for the concentration parameters. Our prior belief about the mode of L_J^*, the number of unique parameter values over holiday destinations, is 15 and about L_D^*, the number of unique parameter values over household compositions, is 5. The prior distributions that match these beliefs are $\lambda_J \sim \text{Gamma}(7.15 \times 20, 20)$ with $\text{var}(L_J^*) = 8.852$ and $\lambda_D \sim \text{Gamma}(3.47 \times 1, 1)$ with $\text{var}(L_D^*) = 3.352$. As in the simulation study in Appendix B, we allow for a wide range of plausible values for the model parameters within a multinomial choice model, and set the prior variance of the model parameters equal to $\sigma_\alpha^2 = \sigma_\beta^2 = 1$.

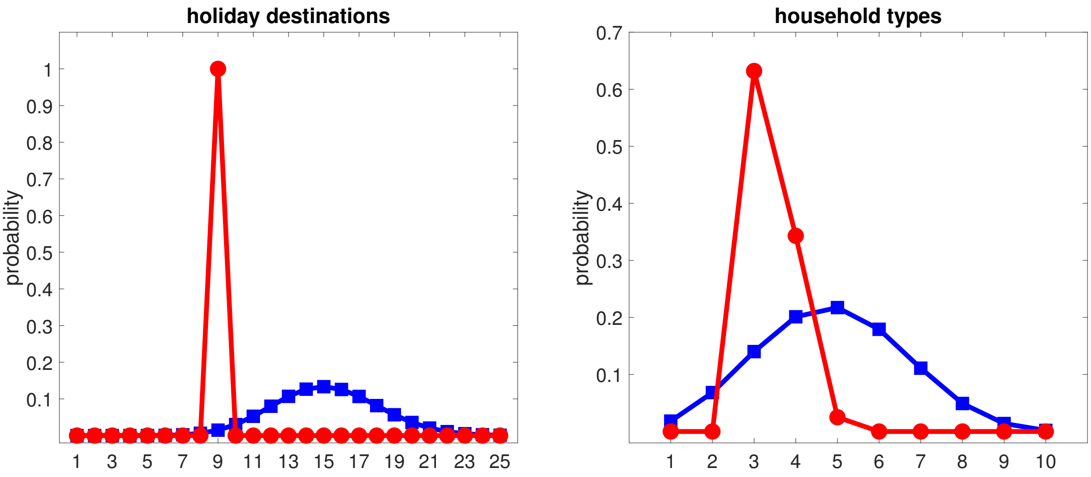
Posterior results are based on 200,000 iterations of the Gibbs sampler, from which the first 100,000 are discarded, and we use a thinning value of 10. Appendix C.3 shows by means of convergence diagnostics that this number of retained draws is sufficient for posterior inference.

4.3 Results

Figure 3 shows that the two-way mixture model reduces the dimensions over the choice categories from 49 holiday destinations to nine unique parameter values, and over the explanatory categories from 11 household compositions to maximum five unique parameter values. The left panel of Figure 3 shows that after convergence all posterior probability mass is concentrated at nine unique parameters for the holiday destinations. The posterior mode in the right panel is located at three clusters of households compositions, with the remaining 34 and 3 percent of probability mass at four and five clusters, respectively. For both dimensions, a large variance of the prior distributions on the concentration parameters is employed, and the posterior distributions over the number of clusters make a considerable move to the left relative to the prior distributions. This observation suggests that the shift to a more parsimonious model is driven by the data, instead of a prior specification encouraging a small number of clusters.

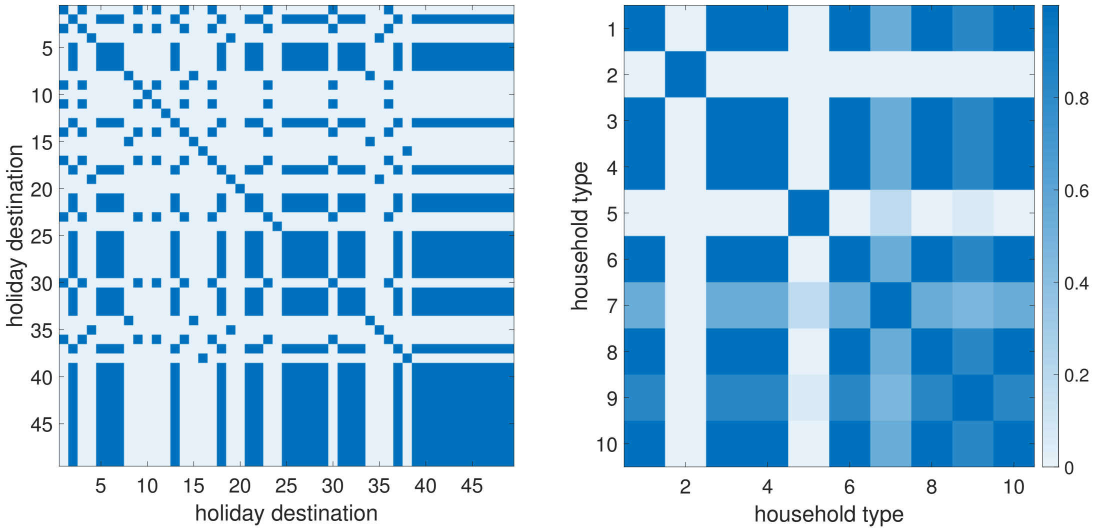
Figure 4 shows which holiday destinations and which household compositions tend to cluster together. The left panel of Figure 4 shows that most holiday

Figure 3: Application: Distribution number of unique parameter values



This figure shows the prior (□ in blue) and posterior (○ in red) distribution over the number of unique parameter values L^* over holiday destinations (left panel) and household compositions (right panel). Appendix A.2 provides more information about the distribution of L^*.

Figure 4: Application: Posterior probabilities cluster memberships

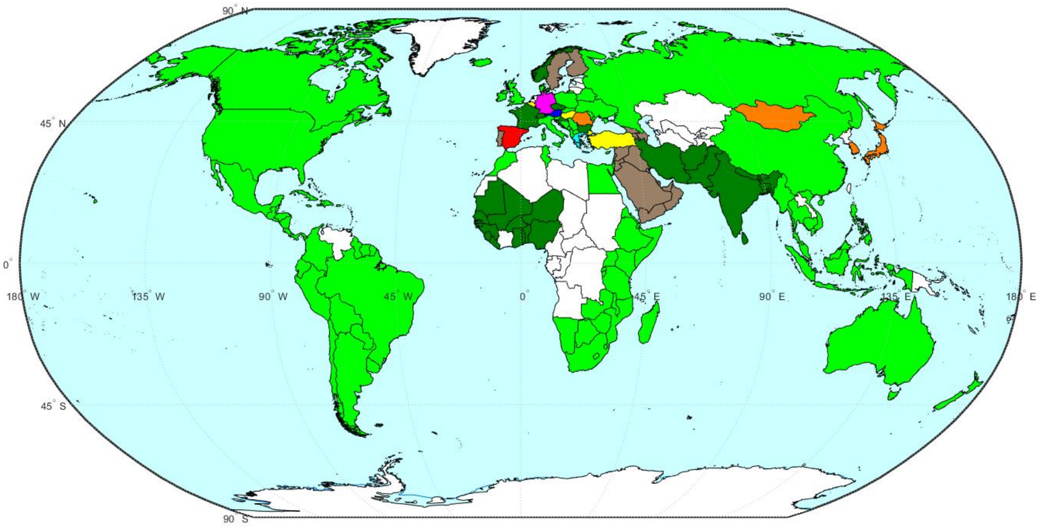


This figure shows the posterior probabilities that the holiday destination at a specific row is in the same cluster as the holiday destination at a specific column (left panel) and the posterior probabilities that household compositions at the rows and columns are in the same cluster (right panel). The posterior probabilities range from zero (light blue) to one (dark blue).

destinations share a cluster with multiple other destinations. The right panel shows the cluster assignment of the household composition dummies. We find that the holiday preferences of single households, in the second household category, and households with children aged 13-17, category five, deviate on average from the other households. The posterior probability that households with children aged 13-17 share the same preferences with households with kids between 0-5 and 13-17 (category 7) is 19 percent, and with household with kids between 0-5, 6-12, and 13-17 (category 9) is 5 percent.

Since the posterior probabilities of cluster memberships of all holiday destinations converge to zero or one, we can infer which destinations share the same clusters (and do not suffer from the label-switching problem). The left panel of Figure 4 shows that there are four categories with their own cluster. That is, for category 10, 12, 20, and 24 no other category than itself has a positive posterior probability of being a cluster member. Figure 1 shows that these categories correspond to Germany, Austria, Spain, and Greece, respectively.

Figure 5: Application: Clustering holiday destinations

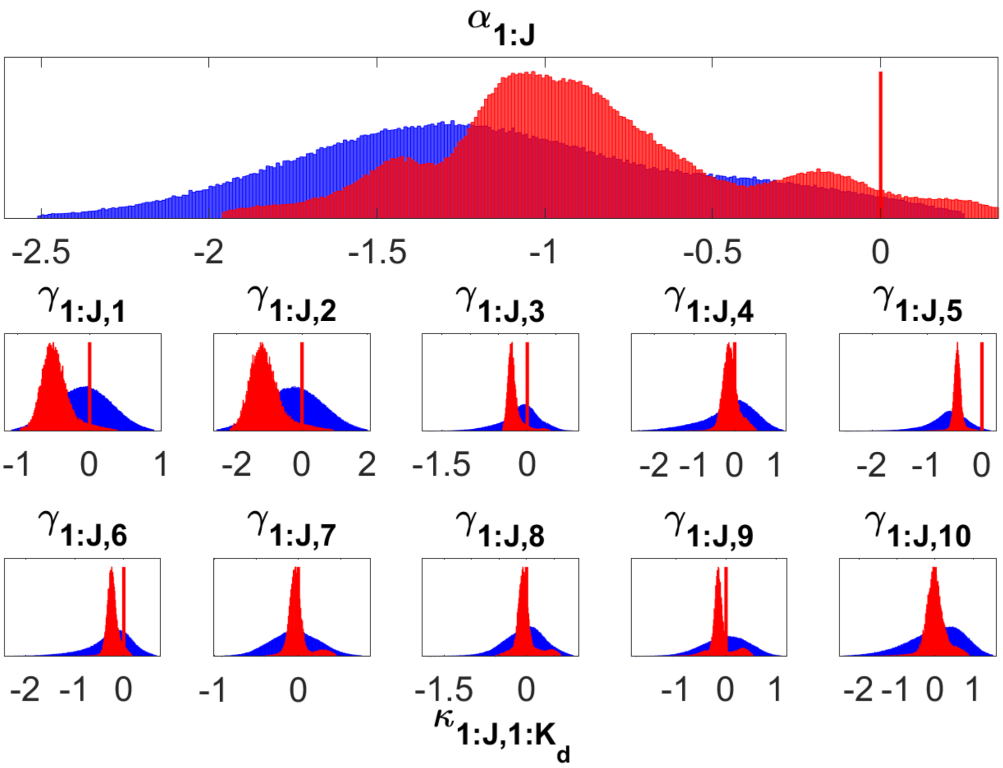


This figure shows the cluster assignments of the holiday destinations of Dutch households. Destinations with the same color are in the same parameter cluster according to Figure 4, and we do not have observations about white regions. The two-way mixture model estimates nine clusters. Appendix C.1 shows the countries in each of the 49 holiday destination choice categories.

Figure 5 shows the parameter clustering over the world's holiday destinations of Dutch households, according to the clustering of the choice categories in the left panel of Figure 4. The sets of destinations with the same color have the same regressor coefficients but different intercepts, which can be interpreted as the base preferences. Conditional on the household characteristics, the households have an expected preference ranking across countries with different colors, but the expected utility only differs in the base preferences between countries with the same color. The estimated parameter clustering is very different from an ad hoc grouping based on, for instance, geographical location.

Comparing the posterior densities of the standard multinomial probit model with the two-way mixture model, we find the densities of the latter model to be more precise. Figure 6 shows the posterior parameter distributions of the explanatory variables over all holiday destinations. The first row shows the parameter distributions of the intercepts, the second and third row of the control variables, and the last row shows the parameter distributions over all explanatory categories

Figure 6: Application: Posterior parameter distributions

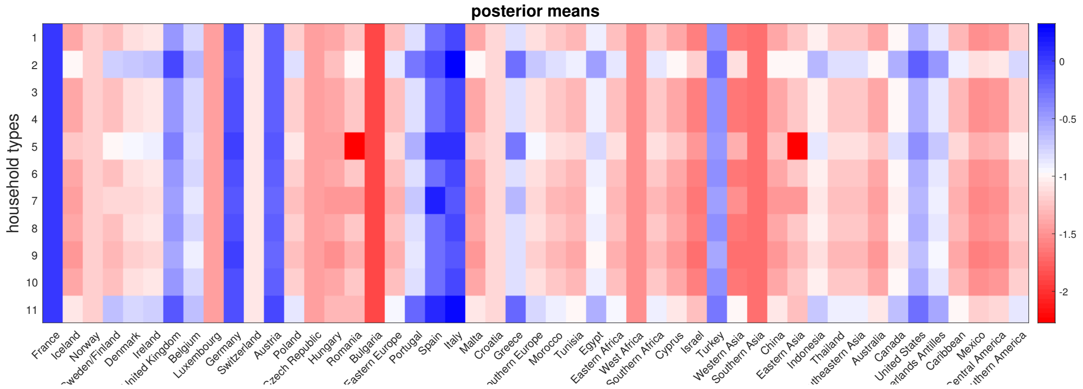


This figure shows the posterior parameter distributions of a standard multinomial probit model (fat in blue) and the two-way mixture model (thin in red). The first row shows the parameter distributions of the intercepts α_j, the second and third row of the K_w control variables γ_{jk}, and the fourth row shows the parameter distributions of κ_{jk} for $j = 1, \ldots, J$ and $k = 1, \ldots, K_d$.

in one window. The mixture model accounts for the uncertainty about the number of clusters, the cluster assignments, and parameter uncertainty. However, sampling separate parameter values for each destination in the standard multinomial choice model results in much more noise. The shapes of the posterior parameter distributions show, except for width, also other differences. Since the mixture model clusters the base category destination France with other destinations, more probability mass is allocated to zero. The posterior distributions of the standard multinomial choice model over all choice categories approximate for most parameters a bell shape. Due to the mixing of parameters over different holiday destinations, the mixture model distributions are more often skewed and show for several parameters multiple modes.

The sum of the household dummy parameters and the intercepts can be interpreted as the average base preferences of each household category. Figure 7 shows these posterior means for each holiday destination and household composition. Since the intercepts are outcome category specific, the preferences vary over each holiday destination. All households have positive base preferences for the traditional popular destinations of Dutch holidaymakers; France, Germany, Austria, Spain, and Italy. However, the second and eleventh household types, households in which every member is 35 or older, have a relatively small difference between base preferences for these popular destinations and other destinations. These households are more inclined to explore countries further away from home.

Figure 7: Application: Posterior parameter means for each category



This figure shows the posterior parameter mean for each household composition category, as showed in Figure 2, over all choice categories in Figure 1. The values of the posterior means are indicated by the color bars with colors ranging from dark red (strongly negative) to dark blue (strongly positive).

Table 1 shows substantial improvements in out-of-sample performance of the two-way mixture model relative to the standard multinomial choice model. The

Table 1: Model evaluation

	two-way		standard		naive	
	in	out	in	out	in	out
hit-rate	0.198	0.186	0.196	0.174	0.149	0.175
likelihood	-15769	-3600	-15274	-3612	-16004	-3614

This table shows in-sample and out-of sample performance for predicting actual category choices measured by hit-rates and log-likelihood as defined in Appendix A.4. The performance of the two-way mixture model 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.

out-of-sample hit-rate of the two-mixture model equals 0.186, where the standard choice model (0.174) does not improve upon the naive prediction method (0.175). The same holds for the out-of-sample likelihood, where the standard model is only slightly better than the naive method, and the two-mixture model shows the best fit. The gains in predictive performance may be explained by Figure 6, which suggests that the two-way mixture model is more efficient in estimating the model parameters than the standard multinomial choice model.

5 Conclusion

With choice data, the number of model parameters typically becomes large. Categorical characteristics of the decision makers enter the model as sets of dummy variables, in which each variable has its own choice alternative specific parameter. The two-way Dirichlet process mixture model clusters parameters over the choice categories and the explanatory dummy categories, while taking the relation between the dependent and independent variables into account.

A high-dimensional empirical application examines how preferences for holiday destinations differ across household types. The estimated clustering over the holiday destinations substantially reduces the parameter uncertainty, resulting in interpretable relations between a large number of household characteristics and a large choice set. The posterior probabilities of households sharing the same preferences allows for targeting a relatively small number of different costumer groups, without imposing a costumer grouping a priori. With potentially more than a thousand parameters, the mixture model finds heterogeneous holiday preferences of a large number of households over a large set of holiday destinations. For instance, we find on average deviating holiday preferences of households with teenagers, and households in which everyone is 35 or older are more inclined to visit holiday destinations far away from home. Moreover, the clustering over outcome categories and explanatory categories results in a substantial gain in predictive performance relative to a standard choice model.

References

Albert, J. H. and Chib, S. Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679, 1993.

Antoniak, C. E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics*, 2(6):1152–1174, 1974.

Bauwens, L., Carpantier, J.-F., and Dufays, A. Autoregressive moving average infinite hidden markov-switching models. *Journal of Business & Economic Statistics*, 35(2):162–182, 2017.

Bernstein, F., Modaresi, S., and Sauré, D. A dynamic clustering approach to data-driven assortment personalization. *Management Science*, 2018.

Bondell, H. D. and Reich, B. J. Simultaneous factor selection and collapsing levels in ANOVA. *Biometrics*, 65(1):169–177, 2009.

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.

Carson, R. T. and Louviere, J. J. Statistical properties of consideration sets. *Journal of Choice Modelling*, 13:37–48, 2014.

Chib, S. and Hamilton, B. H. Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110(1):67–89, 2002.

Chiong, K. X. and Shum, M. Random projection estimation of discrete-choice models with large choice sets. *Management Science*, 2018.

Conley, T. G., Hansen, C. B., McCulloch, R. E., and Rossi, P. E. A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, 144(1):276–305, 2008.

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

Escobar, M. D. and West, M. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Gertheiss, J., Tutz, G., et al. Sparse modeling of categorial explanatory variables. *The Annals of Applied Statistics*, 4(4):2150–2180, 2010.

Geweke, J. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In *Bayesian Statistics*, pages 169–193. University Press, 1992.

Gilbride, T. J. and Allenby, G. M. A choice model with conjunctive, disjunctive, and compensatory screening rules. *Marketing Science*, 23(3):391–406, 2004.

Hirano, K. Semiparametric Bayesian inference in autoregressive panel data models. *Econometrica*, 70(2):781–799, 2002.

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

Hu, X., Munkin, M. K., and Trivedi, P. K. Estimating incentive and selection effects in the medigap insurance market: An application with Dirichlet process mixture model. *Journal of Applied Econometrics*, 30(7):1115–1143, 2015.

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: Smoothing and prior information. *Journal of Computational and Graphical Statistics*, 11(3):508–532, 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.

Kim, S., Shephard, N., and Chib, S. Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65(3): 361–393, 1998.

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.

Newey, W. K. and West, K. D. A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703–708, 1987.

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*, 4: 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.

Tibshirani, R. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Tutz, G. and Gertheiss, J. Regularized regression for categorical data. *Statistical Modelling*, 16(3):161–200, 2016.

Van den Hauwe, S. *Topics in Applied Macroeconometrics.* PhD thesis, Erasmus School of Economics, 2015.

Van Hasselt, M. Bayesian inference in a sample selection model. *Journal of Econometrics*, 165(2):221–232, 2011.

Wiesenfarth, M., Hisgen, C. M., Kneib, T., and Cadarso-Suarez, C. Bayesian nonparametric instrumental variables regression based on penalized splines and Dirichlet process mixtures. *Journal of Business & Economic Statistics*, 32(3): 468–482, 2014.

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

A Bayesian inference

A.1 Truncation level

The stick-breaking representation of the Dirichlet process prior, as in Section 2.3.2, provides a guideline for selecting the truncation level L. When the higher order probabilities $p = \{p_l\}_{l=L}^{\infty}$ in (9), or q for the explanatory categories, are small enough, the approximation error is negligible. Ishwaran and Zarepour (2000) derive the moments of the tail probability $\sum_{l=L}^{\infty} p_l$,

$$\mathrm{E}\left[\sum_{l=L}^{\infty} p_l\right] = \left(\frac{\lambda}{\lambda+1}\right)^{L-1}, \quad \mathrm{var}\left[\sum_{l=L}^{\infty} p_l\right] = \left(\frac{\lambda}{\lambda+2}\right)^{L-1} - \left(\frac{\lambda}{\lambda+1}\right)^{2L-2}, \tag{21}$$

which are the mean and the variance of the tail probability, respectively. Using these statistics, we can test for a particular concentration parameter λ whether the truncation level results in a small enough approximation error.

A.2 Concentration parameter

The concentration parameter λ controls the number of clusters. Hence, the value for λ implies a prior distribution for the number of unique parameter values L^*,

$$Pr[L^* = j|\lambda] = c(j, J)J!\lambda^j \frac{\Gamma(\lambda)}{\Gamma(\lambda + J)}, \tag{22}$$

where we cluster over $j = 1, \ldots, J$ parameters, and $c(j, J) = Pr[L^* = j|\lambda = 1]$. This implied prior distribution over the number of clusters is derived by Antoniak (1974), and Escobar and West (1995) discuss how the factors $c(j, J)$ are calculated. The distribution runs from $L^* = 1$, which means no parameter variation at all, to $L^* = J$ with unique parameter values for each j.

Suppose we have a prior belief about the number of clusters L^*. Van den Hauwe (2015) proposes to choose a value for the concentration parameter λ that sets the prior mode of L^* equal to that belief. The concentration parameter that matches the belief mode$[L^*] = m^*$ is

$$\lambda_{m^*} = \frac{1}{2}\left(\exp(-\delta c(m^*+1)) + \exp(-\delta c(m^*))\right), \tag{23}$$

with $\delta c(1) = \log(c(1,J))$ and $\delta c(m^*) = \log(c(m^*,J)) - \log(c(m^*-1,J))$ for numerical stability.

By choosing λ as in (23) we control the prior mode of the distribution of the number of clusters. Conley et al. (2008) shows that for a range of fixed values for the concentration parameters, the prior distributions on the number of clusters are very informative. By putting a prior on the concentration parameter, we can also govern the variance of the distribution of the number of clusters.

We specify a prior distribution on λ with prior mean equal to the value in (23). To check whether the prior induces enough dispersion around the prior mode of L^*, we evaluate the marginal prior probability density function

$$f(L^*) = \int f(L^*|\lambda) f(\lambda) d\lambda, \tag{24}$$

with Monte Carlo integration, where $f(L^*|\lambda)$ is the probability function in (22) and $f(\lambda)$ is the prior probability density function of λ.

A.3 Posterior simulation

Let $C^* = \{C_1^*, \dots, C_{m_j}^*\}$ denote the current m_j unique values of C excluding $C = 1$, and r_l the number of values in C which equal l. Let $D^* = \{D_1^*, \dots, D_{m_d}^*\}$ denote the current m_d unique values of D. The sampling steps are:

Step 0. Sample the initial draw for the model parameters as $\tilde{\beta}_l|\sigma_\beta \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$. The initial draw for the concentration parameters is $\lambda_J|\eta_{J1}, \eta_{J2} \sim \text{Gamma}(\eta_{J1}, \eta_{J2})$ and $\lambda_D|\eta_{D1}, \eta_{D2} \sim \text{Gamma}(\eta_{D1}, \eta_{D2})$ and for the latent variables $p|\lambda_J \sim \text{stick}(\lambda_J)$, $C_j|p \sim \sum_{l=1}^{L_J} p_l \delta(l)$, $q|\lambda_D \sim \text{stick}(\lambda_D)$, and $D_k|q \sim \sum_{l=1}^{L_D} q_l \delta(l)$. Initialize the latent variables z_i by a draw from a standard normal distribution and center the vector at zero. Permute the elements so that the maximum of each z_i coincides with y_i.

Step 1. Given $\tilde{\beta}$, C, σ_α, z, and x, sample the intercept parameters α according to

$$\alpha_j|\tilde{\beta}, C, \sigma_\alpha, z, x \sim \mathcal{N}(a_j, A^{-1}),\ a_j = A^{-1}\sum_{i=1}^{N} z_{ij} - \tilde{\beta}'_{C_j}x_i,\ A = N + \sigma_\alpha^{-2},$$

for $j = 2, \ldots, J$.

Step 2. Given C, D, σ_β, α, z, and x, sample the model parameters $\tilde{\beta}_l$ for $l = 1, \ldots, L_J$. Distinguish three different cases to sample all parameters in $\tilde{\beta}_l$:

1. For $l \in \{C_1^*, \ldots, C_{m_j}^*\}$ and $\tilde{\kappa}_{l,k}$ with $k \in \{D_1^*, \ldots, D_{m_d}^*\}$, sample $\beta_l^* = (\tilde{\gamma}_l', \tilde{\kappa}_{l,D_1^*}, \ldots, \tilde{\kappa}_{l,D_{m_d}^*})$ in

$$Z_l = \beta_l^* X_l' + \eta, \tag{25}$$

where η is a $1 \times (r_l \times N)$ matrix with independent and identically standard normal distributed elements. The dependent variable $Z_l = (z_1^l, \ldots, z_N^l)$ is defined as a $1 \times (r_l \times N)$ matrix, in which z_i^l are row vectors stacking all $z_{ij} - \alpha_j$ for which $C_j = l$. Aggregate the dummies within each cluster, $x_i^* = (w_i', \sum_{k:D_k=D_1^*} d_{ik}, \ldots, \sum_{k:D_k=D_{m_d}^*} d_{ik})'$, set $x^* = (x_1^*, \ldots, x_N^*)'$, and stack r_l times the matrix x^* in the $(r_l \times N) \times (K_w + m_d)$ matrix X_l. Sample β_l^* according to

$$\beta_l^*|C, D, \sigma_\beta, z, x \sim \mathcal{N}(b, B^{-1}),\ b = Z_l X_l B^{-1},\ B = X_l{}'X_l + \frac{1}{\sigma_\beta^2} I_{K_w+m_d}.$$

2. For $l \in C - \{C_1^*, \ldots, C_{m_j}^*\}$ and $\tilde{\kappa}_{l,k}$ with $k \in \{D_1^*, \ldots, D_{m_d}^*\}$, sample β_l^* from the base distribution as $\beta_l^*|C, D, \sigma_\beta, z, x \sim \mathcal{N}(0, \sigma^2 I_{K_w+m_d})$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$.

3. For $l \in C$ and $\tilde{\kappa}_{l,k}$ with $k \in D - \{1, D_1^*, \ldots, D_{m_d}^*\}$, sample $\tilde{\kappa}_{lk}$ from the base distribution as $\tilde{\kappa}_{lk}|C, D, \sigma_\beta, z, x \sim \mathcal{N}(0, \sigma^2)$, where $\sigma^2 = \sigma_\beta^2$ when $l \neq 1$ and $\sigma^2 = 0$ when $l = 1$.

Step 3. Given p, α, $\tilde{\beta}$, D, z, and x, sample the classification vector of the outcome categories $C = (1, C_2, \dots, C_J)$ according to

$$C_j | p, \alpha, \tilde{\beta}, D, z, x \sim \sum_{l=1}^{L_J} \pi_{lj} \delta_l, \tag{26}$$

for $j = 2, \dots, J$. The conditional cluster probability π_{lj} is a function of the unconditional cluster probability p_l and the likelihood contributions of the latent utilities of each outcome category z_j and the observed explanatory variables x, for the parameter value $\tilde{\beta}_l$. Since $z_{i1}, \dots, z_{iJ}$ are conditionally independent,

$$\begin{aligned}(\pi_{1j}, \dots, \pi_{L_J,j}) \propto & \left(p_1 \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (z_{ij} - \alpha_j - \tilde{\gamma}_1' w_i - \sum_{k=1}^{K_d} \tilde{\kappa}_{1,D_k} d_{ik})^2 \right), \right. \\ & \left. \dots, p_{L_j} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} (z_{ij} - \alpha_j - \tilde{\gamma}_{L_j}' w_i - \sum_{k=1}^{K_d} \tilde{\kappa}_{L_J,D_k} d_{ik})^2 \right) \right).\end{aligned}$$

Step 4. Given q, α, $\tilde{\beta}$, C, z, and x, sample the classification vector of the explanatory categories $D = (D_1, \dots, D_{K_d})$ according to

$$D_k | q, \alpha, \tilde{\beta}, C, z, x \sim \sum_{l=1}^{L_D} \psi_{lk} \delta_l, \tag{27}$$

for $k = 1, \dots, K_d$. Since clusters of different explanatory dummies 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

$$\ddot{\kappa}_{C_j,kl} = (\tilde{\kappa}_{C_j,D_1}, \dots, \tilde{\kappa}_{C_j,D_{k-1}}, \tilde{\kappa}_l, \tilde{\kappa}_{C_j,D_{k+1}}, \dots, \tilde{\kappa}_{C_j,D_{K_d}}),$$

which is the coefficient vector $\tilde{\kappa}_{C_j}$ based on the classification vector D 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 conditional cluster probabilities ψ_{lk} are a function of the unconditional cluster probabilities q_l and the data (z, x),

$$(\psi_{1k}, \ldots, \psi_{L_D,k}) \propto \left(q_1 \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{J} (z_{ij} - \alpha_j - \tilde{\gamma}'_{C_j} w_i - \ddot{\kappa}_{C_j,k1} d_i)^2 \right), \right.$$
$$\left. \ldots, q_{L_d} \exp\left(-\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{J} (z_{ij} - \alpha_j - \tilde{\gamma}'_{C_j} w_i - \ddot{\kappa}_{C_j,k,L_D} d_i)^2 \right) \right).$$

Step 5. Given C and λ_J, sample the unconditional cluster probabilities for the outcome categories from $p|C, \lambda_J$ according to

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

where

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

Given D and λ_D, sample the unconditional cluster probabilities for the explanatory categories q in the same way as for p.

Step 6. Given p, η_{J1}, and η_{J2}, sample the concentration parameter for the outcome categories λ_J according to

$$\lambda_J | p, \eta_{J1}, \eta_{J2} \sim \text{Gamma}\left(L_J + \eta_{J1} - 1, \eta_{J2} - \sum_{l=1}^{L_J - 1} \log(1 - V_l^*) \right). \tag{28}$$

Given q, η_{D1}, and η_{D2}, sample the concentration parameter for the explanatory categories λ_D in the same way as for λ_J.

Step 7. Given α, $\tilde{\beta}$, C, D, y, and x, sample the latent variables z_{ij} for $i = 1, \ldots, N$ and for $j = 1, \ldots, J$. Following from (1), $z_{ij} \geq \max(z_i^{(j)})$ if $y_i = j$ and $z_{ij} \leq \max(z_i^{(j)})$ if $y_i \neq j$, where $z_i^{(j)} = (z_{i1}, \ldots, z_{i,j-1}, z_{i,j+1}, \ldots, z_{iJ})$. Hence, sample z_{ij} for $j = y_i$ and for $j \neq y_i$ respectively according to

$$z_{ij}|z_i^{(j)}, \alpha_j, \tilde{\beta}_{C_j}, D, y_i, x_i \sim \mathcal{N}_{+\max(z_i^{(j)})}\left(\alpha_j + \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik}, 1\right),$$
$$z_{ij}|z_i^{(j)}, \alpha_j, \tilde{\beta}_{C_j}, D, y_i, x_i \sim \mathcal{N}_{-\max(z_i^{(j)})}\left(\alpha_j + \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik}, 1\right),$$

where $\mathcal{N}_{+a}(\mu, \sigma^2)$ and $\mathcal{N}_{-a}(\mu, \sigma^2)$ represent a normal distribution with expectation μ and variance σ^2 truncated from below or above by a, respectively.

Step 8. Go to Step 1.

Note that this sample algorithm clusters parameters over both outcome and explanatory categories. In case we only want to cluster over outcome categories, we simply put all explanatory variables in w_i. The vector d_i remains empty, which means that we do not have to restructure the dummy variables and sample their parameters $\tilde{\kappa}$ in Step 2, and ignore Step 4 of the sample algorithm. On the other hand, when we only cluster parameters over explanatory variables, we set $L_J = J$, $C = (1, 2, \ldots, J)$, and skip Step 3.

A.4 Predictive distributions and evaluation criteria

The predictive densities of y_i for different individuals $i = 1, \ldots, N$ are simulated by means of (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

$$y_i^{(s)}(z_{ij}^{(s)}) = j \text{ if } z_{ij}^{(s)} = \max(z_i^{(s)}), \tag{29}$$

where $z_i^{(s)} = (z_{i1}^{(s)}, \ldots, z_{iJ}^{(s)})$ is a vector with draws of the latent utilities in iteration s of the sampler. We obtain a draw from the predictive density of z_{ij} as follows

$$z_{ij}^{(s)} = \alpha_j^{(s)} + \tilde{\gamma}_{C_j^{(s)}}^{(s)'} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j^{(s)},D_k^{(s)}}^{(s)} d_{ik} + \varepsilon_{ij}^{(s)}, \quad \varepsilon_{ij}^{(s)} \sim \mathcal{N}(0, I_J), \tag{30}$$

for $j = 1, \ldots, J$, where $\alpha_j^{(s)}$, $\tilde{\gamma}^{(s)}$ and $\tilde{\kappa}^{(s)}$ are the model parameter draws for α_j, $\tilde{\gamma}$ and $\tilde{\kappa}$, and $C_j^{(s)}$ and $D_k^{(s)}$ are the classification parameter draws for C_j and D_k in iteration s of the sampler.

The predictive distribution provides the estimated choice probabilities by

$$\hat{P}(y_i = j) = \frac{1}{S} \sum_{s=1}^{S} I[y_i^{(s)} = j], \tag{31}$$

where S denotes the number of samples from the predictive density, $I(A)$ is an indicator function that equals one if event A occurs and zero otherwise, and $y_i^{(s)}$ is defined in (29).

The accuracy of the predictive density is measured by the hit-rate and the log-likelihood. The hit-rate is defined as

$$\text{HR} = \frac{1}{NJ} \sum_{i=1}^{N} \sum_{j=1}^{J} I[\hat{P}(y_i = j) \geq \hat{P}(y_i = k) \forall k] I[y_i = j], \tag{32}$$

and the log-likelihood equals

$$\text{LL} = \sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \ln(\hat{P}(y_i = j)) + (1 - I[y_i = j]) \ln(1 - \hat{P}(y_i = j)). \tag{33}$$

B Numerical experiment

This appendix examines the practical implications of the parameter clustering methods on simulated data. We estimate the two-way mixture model and compare the performance to a standard multinomial choice model.

B.1 General set-up

The choice data are generated from a multinomial choice model with control variables and a categorical explanatory variable. The outcome categories and the explanatory categories vary both over two parameter clusters. The data generating process takes the form

$$
\begin{aligned}
y_i(z_i) &= j \text{ if } z_{ij} = \max(z_i), \\
z_{ij} =& \alpha_j + \gamma_j' w_i + \kappa_j' d_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathcal{N}(0,1),
\end{aligned}
\tag{34}
$$

with $j = 1, \ldots, J$ and $i = 1, \ldots, N$. The vector $w_i = (w_{i1}, w_{i2})$ includes two standard normally distributed variables $w_i \sim \mathcal{N}(0, I_2)$. The categorical dummies are drawn from a multinomial distribution

$$
(d_{i1}, \ldots, d_{i,K_d}, d_{i,K_d+1}) \sim \text{Multinomial}\left(\frac{p_{d_i}}{K_d}, \ldots, \frac{p_{d_i}}{K_d}, 1 - p_{d_i}\right), \tag{35}
$$

where $p_{d_i} = \frac{\exp(w_{i2})}{1+\exp(w_{i2})}$ and $d_i = (d_{i1}, \ldots, d_{i,K_d})$.

We mimick the dimensions of the empirical application in Section 4 and apply the Gibbs sampler to $N = 4000$ observations simulated from the data generating process, and use another 1000 observations for out-of-sample analysis. We set the number of outcome categories to $J = 50$ and the number of explanatory categories to $K_d = 10$. The outcome and explanatory categories are clustered into two groups, with model parameter values $\tilde{\beta}_l = (\tilde{\gamma}_l', \tilde{\kappa}_{l,1}, \ldots, \tilde{\kappa}_{l,L_D})$ equal to

$$
\begin{aligned}
\tilde{\beta}_1 &= (\tilde{\gamma}_1', \tilde{\kappa}_{1,1}, \tilde{\kappa}_{1,2}) = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}, \\
\tilde{\beta}_2 &= (\tilde{\gamma}_2', \tilde{\kappa}_{2,1}, \tilde{\kappa}_{2,2}) = \begin{pmatrix} -1 & 1 & 0 & 2 \end{pmatrix},
\end{aligned}
\tag{36}
$$

where $\beta_j = (\tilde{\gamma}_{C_j}', \tilde{\kappa}_{C_j,D_1}, \ldots, \tilde{\kappa}_{C_j,D_{10}})$ with $C_j = 1$ for $j = 1, \ldots, 25$ and $C_j = 2$ for $j = 26, \ldots, 50$, and $D_k = 1$ for $k = 1, \ldots, 5$ and $D_k = 2$ for $k = 6, \ldots, 10$. The intercepts $\alpha_j = 0$ for $j = 1, \ldots, 25$ and $\alpha_j = -1$ for $j = 26, \ldots, 50$.

We set $L_J = L_D = 10$. Since $L_D = K_d$, we estimate a full Dirichlet process for the explanatory categories. By truncating the number of possible outcome category clusters to L_J, we obtain a potential approximation error. Following the procedure in Appendix A.1, the expectation and the variance of the aggregated higher order probabilities equal 0.001 and 2.979×10^{-5} for the sampled α_J in the last iteration of the Gibbs sampler for the two-way mixture model. These numbers confirm that the approximation error is negligible in the posterior simulation.

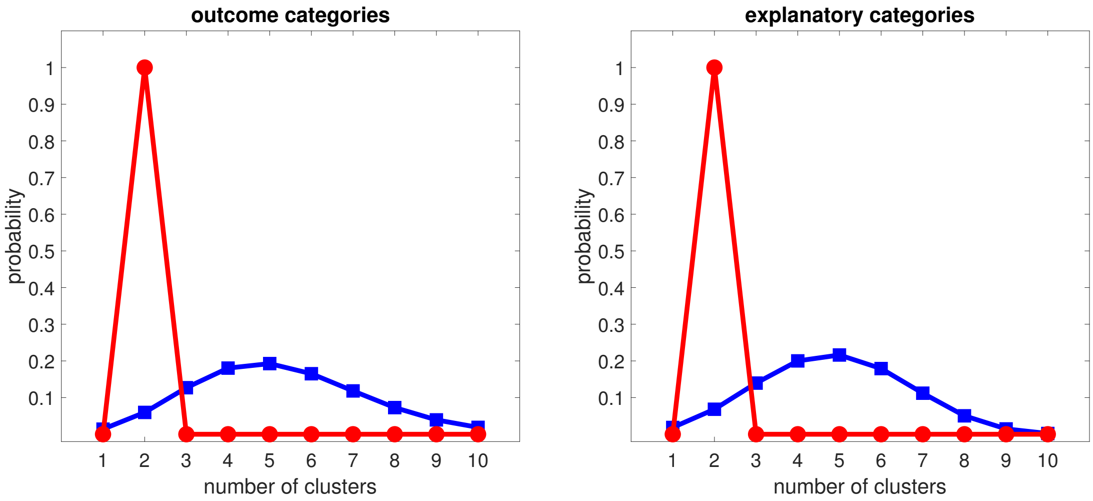
The priors for the concentration parameters are parametrized according to the procedure in Section A.2. The means of the prior distributions equal the concentration parameters that match the prior belief that the mode of unique parameter values equals five. That results in the distributions $\alpha_J \sim \text{Gamma}(1.30 \times 10, 10)$ with $\text{var}(L_J^*) = 4.27$ and $\alpha_D \sim \text{Gamma}(3.47 \times 1, 1)$ with $\text{var}(L_D^*) = 2.99$. The prior variance of the model parameters is set to $\sigma_\alpha^2 = \sigma_\beta^2 = 1$, which allows for a wide range of plausible values within a multinomial choice model.

Posterior results are based on 200,000 iterations of the Gibbs sampler, from which the first 100,000 are discarded and we use a thinning value of 10.

B.2 Results

The mixture model correctly identifies the number of different parameter clusters and cluster memberships of the categories. Figure 8 shows the posterior distributions together with the prior distributions for the number of distinct parameter values for both the outcome categories and the explanatory categories. The posteriors for both dimensions shift all probability mass under the relatively uninformative prior to two clusters, which is equal to the number of clusters in the data

Figure 8: Two-way: Distribution number of unique parameter values



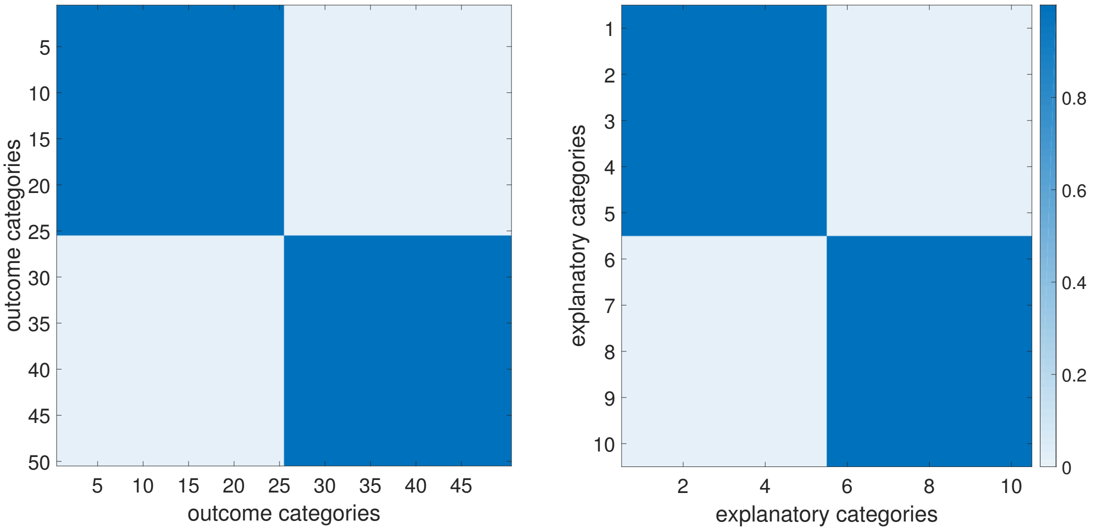
This figure shows the prior (□ in blue) and posterior (○ in red) distribution over the number of unique parameter values L^* over the outcome categories (left panel) and the explanatory categories (right panel) in the two-way mixture model.

generating process.

Figure 9 displays that the cluster assignment of the categories equal to what is expected based on the data generating process. With posterior probability equal to zero, two categories share a parameter cluster when that is not the case in the data generating process. All first 25 outcome categories are assigned to a cluster with the base category, in which all parameter values are exactly zero, and also the final 25 categories are in one and the same cluster with probability one. The same findings hold for the explanatory categories.

Mixing a large number of categories into a relatively small number of parameter clusters improves in parameter estimation efficiency. Figure 10 shows the posterior parameter distributions of the two-way mixture model and a standard multinomial probit model. The mixture model estimates considerable thinner posterior distributions compared to the standard multinomial choice model. Moreover, the posterior parameter distributions of the mixture model are centered around the

Figure 9: Two-way: Posterior probabilities cluster memberships



This figure shows the posterior probabilities that the outcome category at a specific row is in the same cluster as the outcome category at a specific column (left panel) and the posterior probabilities that explanatory categories at the rows and columns are in the same cluster (right panel) in the two-way clustering model. The posterior probabilities range from zero (light blue) to one (dark blue).

Table 2: Diagnostics posterior parameter distributions

			cluster			standard		
j		DGP	MSE	MAE	IQR	MSE	MAE	IQR
Jan-25	α_j	0.000	0.012	0.090	0.087	0.016	0.101	0.115
	γ_1	0.000	0.000	0.000	0.000	0.011	0.083	0.098
	γ_2	0.000	0.000	0.000	0.000	0.018	0.108	0.103
	$\bar{\kappa}_{1:5}$	0.000	0.000	0.000	0.000	0.126	0.283	0.350
	$\bar{\kappa}_{6:10}$	0.000	0.000	0.000	0.000	0.800	0.723	0.532
26-50	α_j	-1.000	0.021	0.122	0.110	0.094	0.267	0.180
	γ_1	-1.000	0.003	0.042	0.047	0.059	0.218	0.120
	γ_2	1.000	0.002	0.035	0.052	0.050	0.198	0.128
	$\bar{\kappa}_{1:5}$	0.000	0.015	0.109	0.080	0.294	0.397	0.479
	$\bar{\kappa}_{6:10}$	2.000	0.031	0.160	0.096	1.356	1.111	0.350

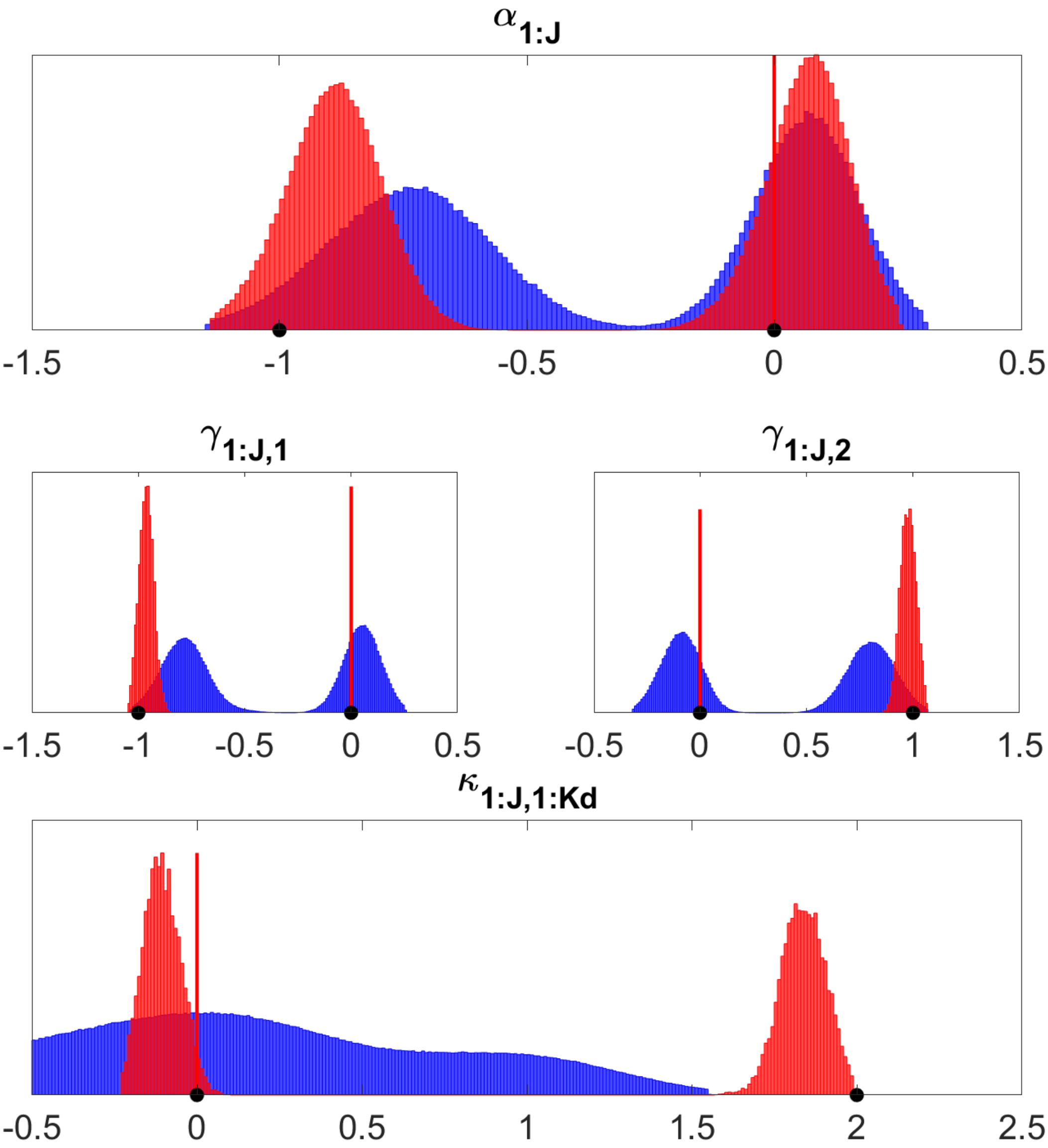
This table shows the performance measures for the parameters averaged over the outcome categories in the first cluster in the first five rows, and for the second cluster in the last five rows. The parameter draws for the dummies in the same cluster are averaged in $\bar{\kappa}_{j,1:5} = \frac{1}{5}\sum_{k=1}^{k=5}\kappa_{jk}$ and $\bar{\kappa}_{j,6:K_d} = \frac{1}{5}\sum_{k=6}^{k=K_d}\kappa_{jk}$. The first column shows the parameter values in the data generating process, the next column the mean squared error, the mean absolute error, and the interquartile range.

parameter values in the data generating process, where the posterior modes of the standard model deviate from these values. Note that outcome categories clustered with the first outcome category have parameter values exactly equal to zero, resulting to an accumulation of probability mass at zero in the posterior distributions of the mixture model.

The posterior parameter distribution diagnostics in Table 2 formalize the gains in estimation performance due to parameter clustering. The mixture model outperforms the benchmark for all diagnostics on each estimated parameter. The posterior parameter means of the mixture model are much closer to the parameter values in the data generating process, which is confirmed by the values of the mean squared error and the mean absolute error of the parameter draws. The relative interquartile range values show that the Dirichlet process prior is more efficient in exploiting sample information than the standard multinomial choice model.

Table 3 shows that two-way parameter clustering also yields higher out-of-

Figure 10: Posterior parameter distributions two-way clustering



This figure shows the posterior parameter distributions of a standard multinomial probit model (fat, in blue) and the two-way clustering model (thin, in red). The first row shows the parameter distributions of the intercepts α_j, the second row of the K_w control variables γ_{jk}, and the third row shows the parameter distributions of κ_{jk} for $j = 1, \ldots, J$ and $k = 1, \ldots, K_d$. The black dots represent the parameter values in the data generating process in (36).

sample hit rates and likelihoods relative to a standard choice model. In-sample,

Table 3: Diagnostics in-sample and out-of-sample model fit

	two-way		standard		naive	
	in	out	in	out	in	out
hit-rate	0.044	0.041	0.059	0.025	0.031	0.031
likelihood	-17755	-4434	-17515	-4510	-19437	-4879

This table shows in-sample and out-of sample performance for predicting actual category choices measured by hit rates and log-likelihood as defined in Appendix A.4. The performance of the two-way mixture model 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.

the standard model achieves a better hit rate, probably due to the large number of parameters. However, this gain in accuracy comes at the cost of efficiency. Therefore, the out-of-sample hit rate is better for the mixture model.

C Empirical application

C.1 Overview categorical dependent variable

This appendix shows the countries within each holiday destination choice category in Figure 1 in Section 4.

Eastern Europe
Belarus
Moldova
Ukraine
Slovakia
Russia
Southern Europe
Slovenia
Albania
Bosnia and Herzegovina
Gibraltar
Vatican City
Montenegro
San Marino
Serbia
Macedonia
Eastern Africa
Kenya
Burundi
Comoros
Djibouti
Eritrea
Ethiopia
Madagascar
Malawi
Mauritius
Mayotte
Mozambique
Reunion
Rwanda
Seychelles
Somalia
Uganda
Tanzania
Zambia
Zimbabwe
West Africa
Gambia
Benin
Burkina Faso
Cape Verde
Cote dIvoire
Ghana
Guinea
Guinea-Bissau
Liberia
Mali
Mauritania
Niger
Nigeria
Saint Helena
Senegal
Sierra Leone
Togo
Southern Africa
South Africa
Botswana
Lesotho
Namibia
Swaziland
Western Asia
Jordan
Armenia
Azerbaijan
Bahrain
Georgia
Iraq
Kuwait
Lebanon
Oman
Palestine
Qatar
Saudi Arabia
Syrian
United Arab Emirates
Yemen
Southern Asia
Afghanistan
Bangladesh
Bhutan
Iran
Maldives
Nepal
Pakistan
India
Sri Lanka
Eastern Asia
Hong Kong
Japan
Korea
Macau
Mongolia
Southeastern Asia
Brunei
Burma
Cambodia
Laos
Philippines
Singapore
Timor-Leste
Viet Nam
Malaysia
Caribbean
Anguilla
Antigua and Barbuda
Aruba
Bahamas
Barbados
British Virgin Islands
Cayman Islands
Cuba
Dominica
Grenada
Guadeloupe
Haiti
Jamaica
Martinique
Montserrat
Puerto Rico
Saint Barthelemy
Saint Kitts and Nevis
Saint Lucia
Saint Martin
Saint Vincent and the Grenadines
Trinidad and Tobago
Turks and Caicos Islands
United States Virgin Islands
Central America
Belize
Costa Rica
El Salvador
Guatemala
Honduras
Mexico
Nicaragua
Panama
Southern America
Brazil
Argentina
Bolivia
Chile
Colombia
Ecuador
Falkland Islands
French Guiana
Guyana
Paraguay
Peru
Suriname
Uruguay

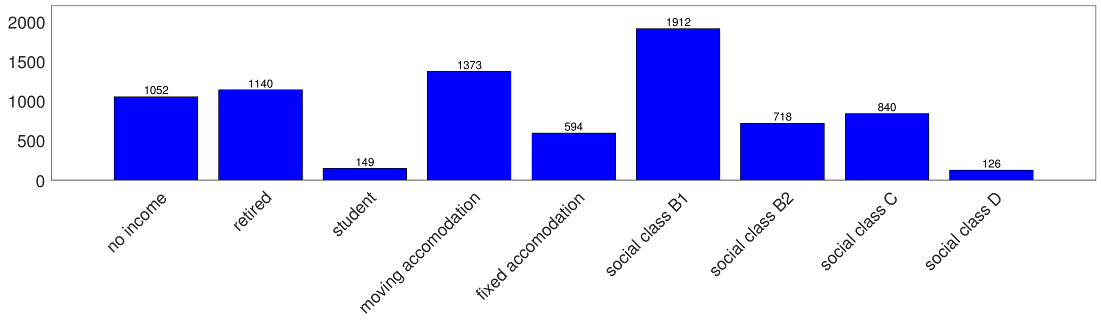
C.2 Overview control variables

Table 4: 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 in the models as the standardized log mean of each income group. We correct for the no responses by including a dummy in the model.

Figure 11: Frequency counts dummy control variables



This figure shows the frequency counts for the explanatory control variables. Moving holiday accommodations include tents, caravans, campers, and cabin boats. Fixed holiday accommodations are defined as holiday homes or a mobile home with a fixed location. The sample is divided in five social classes, captured by four dummy variables. The upper social class A is the reference category, B and C represent the middle class, and D is the lower social class.

C.3 Convergence diagnostics

To infer whether we use enough draws from our posterior simulator, we analyze inefficiency factors $1+2\sum_{f=1}^{\infty}\rho_f$, where ρ_f is the fth order autocorrelation of the chain of draws for a specific parameter. We use the Bartlett kernel as in Newey and West (1987) with a bandwidth of four percent of the sample draws. The inefficiency factors equal the variance of the mean of the posterior draws from the sampler, divided by the variance of the mean assuming independent draws. When we require the variance of the mean of the posterior draws to be limited to at most one percent of the variation due to the data, the inefficiency factor provides an indication of the minimum number of draws to achieve this, see Kim et al. (1998).

We also test for convergence of the sampler by the Geweke (1992) t-test for the null hypothesis of equality of the means computed from the first 20 percent and the last 40 percent of the sample draws. We compute the variances of the means using the Newey and West (1987) heteroskedasticity and autocorrelation robust variance estimator with a bandwidth of four percent of the sample sizes.

Table 5: Summary of simulation convergence tests and inefficiency factors

	Convergence test			Inefficiency factors		
	10%	5%	1%	Mean	Min	Max
control variables						
income	0.042	0.021	0.000	85.971	17.291	106.517
income dummy	0.083	0.021	0.000	85.832	16.567	105.839
retired	0.000	0.000	0.000	17.888	5.523	20.486
student	0.708	0.104	0.000	10.153	2.723	33.337
moving accomodation	0.583	0.583	0.000	10.389	2.790	30.670
fixed accomodation	0.000	0.000	0.000	11.305	3.288	22.115
social class B1	0.021	0.021	0.021	22.858	6.570	26.468
social class B2	0.021	0.000	0.000	23.875	5.392	28.423
social class C	0.021	0.000	0.000	27.655	8.738	31.199
social class D	0.021	0.021	0.000	19.481	4.115	23.503
household categories						
single <35	0.000	0.000	0.000	19.317	4.983	23.033
single >=35	0.000	0.000	0.000	18.086	5.031	21.536
kids 0-5	0.000	0.000	0.000	19.317	4.983	23.033
kids 6-12	0.000	0.000	0.000	19.317	4.983	23.033
kids 13-17	0.021	0.000	0.000	15.589	4.464	60.708
kids 0-5, 6-12	0.000	0.000	0.000	19.317	4.983	23.033
kids 0-5, 13-17	0.083	0.000	0.000	10.226	2.018	16.154
kids 6-12, 13-17	0.000	0.000	0.000	19.317	4.983	23.033
kids 0-5, 6-12, 13-17	0.833	0.792	0.688	181.969	40.255	220.470
hh>1, head<35	0.000	0.000	0.000	19.317	4.983	23.033

This table shows the percentage rejections per significance level on the convergence tests, and statistics of the inefficiency factors, over draws for all outcome categories. Parameters for which all draws are equal to the base category, which parameter values are identical to zero, are not included in this analysis. Note that the seventh and ninth household category only contain three and fifteen in-sample observations, respectively.