MONASH University

MONASH BUSINESS SCHOOL

DEPARTMENT OF ECONOMETRCS AND BUSINESS STATISTICS

ISSN 1440-771X

WORKING PAPER SERIES

A high-dimensional multinomial logit

Didier Nibbering

Working Paper No. 19/23
Nov 2023

A high-dimensional multinomial logit model

Didier Nibbering*

Department of Econometrics and Business Statistics, Monash University

Abstract

The number of parameters in a standard multinomial logit 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 in a multinomial logit model is often large. This paper proposes a new method for data-driven two-way parameter clustering over outcome categories and explanatory dummy categories in a multinomial logit model. A Bayesian Dirichlet process mixture model encourages parameters to cluster over the categories, which reduces the number of unique model parameters and provides interpretable clusters of categories. In an empirical application, we estimate the holiday preferences of 11 household types over 49 holiday destinations, and identify a small number of household segments with different preferences across clusters of holiday destinations.

Keywords: large choice sets, Dirichlet process prior, multinomial logit 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, Bruno Jacobs, Ruben Loaiza-Maya, David Frazier, 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. Since the parameters in a multinomial logit model 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 a dummy variable for each category. With several categorical variables and large numbers of categories, the number of parameters required to describe the effect on one choice alternative is already large.

Due to the structure of the multinomial logit model, reducing the number of parameters is not straightforward. Many regularization methods have been proposed that shrink parameters to zero (Bhadra et al., 2019). Since each explanatory variable in a multinomial logit model has multiple alternative-specific parameters, zero values for a subset of these parameters does not remove the effect of the variable (Vincent and Hansen, 2014). While a zero parameter is equivalent to a zero marginal effect in a linear regression model, shrinking a parameter to zero in the nonlinear multinomial logit model may decrease or even increase the marginal effect on different choice alternatives, and therefore lacks interpretation.

The main contribution of this paper is a Bayesian method that clusters parameters in a multinomial logit model in a data-driven way. In contrast to shrinking parameters to zero, clustering parameters across categories has the unambiguous interpretation that the marginal effect of the explanatory category on the outcome category is the same for clustered categories. We propose a two-way Dirichlet process prior on the model parameters that encourages the alternative-specific parameters to cluster over outcome and explanatory categories. We set up a Gibbs sampler that draws cluster assignments for both dependent and independent categorical variables. The posterior parameter distributions incorporate parameter uncertainty together with uncertainty in cluster assignments and the number of clusters. This cluster uncertainty is ignored when the number of clusters is fixed or when categories are aggregated to higher levels a priori.

The proposed mixture model allows for two-way and one-way clustering across either outcome or explanatory categories. The model is, to the best of our knowledge, the first two-way mixture choice model. Our one-way mixture models build upon recent developments in shrinkage priors. MacLehose and Dunson (2010) propose a Dirichlet process prior that shrinks parameters across explanatory variables to multiple values, extending priors that only shrink to zero. Burgette et al. (2013) implement an adapted version of this prior to shrink parameters across outcome categories in a multinomial probit model with i.i.d. errors. This prior shrinks parameters to a number of locations, but does not estimate clusters of identical parameter values. Moreover, the prior also shrinks the intercepts, which we exclude from the parameter clustering so that they can take the variation across outcome category frequencies into account. Pauger et al. (2019) use a spike and slab prior to cluster explanatory categories in a linear regression model. Their sampler has to iterate over all pairwise differences in each step, instead of iterating over a small number of clusters with a Dirichlet process prior.

The contribution of the proposed method in applied statistics is illustrated with an empirical application in marketing. We estimate the holiday preferences of 11 household types over 49 holiday destinations using survey data. Among holiday destinations with a high cluster probability, all households have similar preferences. For instance, the estimated cluster probability between Turkey and Greece equals 97.9%, and between the Netherlands Antilles and the United States 92.8%. For these destinations, marketing efforts do not have to distinguish different household segments. Between holiday destinations with low cluster probabilities, the holiday preferences vary across households. The cluster probabilities of the household types show among which segments the preferences vary the most. The results suggest that targeting three segments of households –younger than 35 without kids, 35 or older without kids or with teenagers, and households with kids younger than 13– may be more cost effective than targeting each individual household.

To allow for correlation in unobserved factors across choice alternatives, the multinomial logit model is often extended with random coefficients to a mixed logit model. This model can be estimated with alternative-specific covariates and repeated observations as

for example in Train (2009), Greene and Hensher (2010), and Fiebig et al. (2010). Since the empirical application includes household-specific covariates in a cross-sectional data set, Nibbering (2023) specifies a mixed logit model with a two-way Dirichlet process prior, derives a Gibbs sampler, and estimates this model on simulated data.

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. For instance, if the cluster dimension is known a priori, a hierarchical prior can be used that specifies which categories are similar ex ante while allowing the data to determine the degree of similarity ex post (Geweke et al., 2003). Regularization techniques for high-dimensional regressor matrices, such as the lasso introduced by Tibshirani (1996), can be adapted for settings in which the cluster dimension is not known a priori. 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 outline of this paper is as follows. Section 2 discusses the model specification and Section 3 Bayesian inference. Section 4 discusses the empirical application and Section 5 concludes. Appendix A provides an overview of the notation used in this paper.

2 Model specification

2.1 Multinomial logit model

Let y_i be an observable 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. Define $y = (y_1, \ldots, y_N)'$ and $X = (x_1, \ldots, x_N)'$. The probability that individual i chooses alternative j is

$$P(y_i = j|x_i) = \frac{\exp(\eta_{ij})}{\sum_{j=1}^{J} \exp(\eta_{ij})}, \tag{1}$$

where η_{ij} is a linear function of parameters for all $j = 1, \ldots, J$,

$$\eta_{ij} = \alpha_j + x_i'\beta_j, \tag{2}$$

with alternative-specific intercept α_j and K-dimensional coefficient vector $\beta_j = (\beta_{j1}, \ldots, \beta_{jK})'$.

The model parameters $\alpha = (\alpha_1, \ldots, \alpha_J)'$ and $\beta = (\beta_1, \ldots, \beta_J)'$ in (2) are not identified. For any scalar c and for any value of the parameters α_j and β_{jk}, the parameters $\alpha_j + c$ and $\beta_{jk} + c$ result in identical probabilities. To overcome this additive redundancy we set $\alpha_1 = 0$ and $\beta_{1k} = 0$ for all k.

With a large choice set, the number of parameters in the $J \times K$ matrix β is large. Large numbers of parameters amplify overfitting concerns and make it difficult to extract useful insights. For the data to be informative on the parameters without additional restrictions, the number of outcome categories and explanatory variables need to be relatively small. Formulating the multinomial logit model as a set of binary logit models does not change the total number of parameters (Agresti, 2003). The estimates from the binary logit specifications may be less efficient, although the difference is small when the response category having highest prevalence is the baseline (Becg and Gray, 1984).

Two features of many large scale empirical applications of choice models exacerbate

the curse of dimensionality. First, the observed choices in y 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.

2.2 Parameter clustering over categories

When categories are allowed to have the same parameter value, the number of unique parameter values corresponding to a categorical variable can be decreased. This section defines clusters of categories with identical parameters.

2.2.1 Parameter clustering over explanatory categories

To cluster over categories within a categorical explanatory variable, we make an explicit distinction in the vector $x_i = (w_i', d_i')'$. The vector w_i contains K_w explanatory variables for which we do not cluster the parameters. We allow for parameter clustering across the dummy variables within the K_d-dimensional vector d_i. The K_d dummies in $d_i = (d_{i1}, \ldots, d_{iK_d})'$ correspond to the first K_d of the $K_d + 1$ categories in a categorical explanatory variable, and we define K_d+1 as the baseline category. We rewrite the model in (2) to

$$\eta_i = \alpha + \gamma w_i + \kappa d_i + \varepsilon_i = \alpha + \gamma w_i + \sum_{k=1}^{K_d} \kappa_{\cdot k} d_{ik}, \tag{3}$$

where $\eta_i = (\eta_{i1}, \ldots, \eta_{iJ})'$ and $\beta = (\gamma, \kappa)$ with $\kappa = (\kappa_{\cdot 1}, \ldots, \kappa_{\cdot K_d})$. The parameter values in $\kappa_{\cdot k} = (\kappa_{1k}, \ldots, \kappa_{Jk})'$ correspond to the dummy variable d_{ik}, with $k = 1, \ldots, K_d$. We cluster the parameters $\kappa_{\cdot k}$ across the categories within one categorical explanatory variable d_i in (3). Appendix B extends the model with parameter clusters for multiple

explanatory categorical variables.

The parameters in κ vary across dummy categories. Equivalently, we can say that the parameters vary over a number of clusters, where the number of clusters equals the number of categories if each category has a different parameter value. Within the clusters the parameters are assumed to be identical, but across clusters the parameters are allowed to be different. The cluster representation of (3) is

$$\eta_i = \alpha + \gamma w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{.D_k} d_{ik}, \tag{4}$$

where $\kappa_{.k} = \tilde{\kappa}_{.D_k}$ can vary over L_K clusters. The classification variables $D_k \in \{1, \ldots, L_K\}$ take integer values indicating the cluster for category k in the explanatory categorical variable, and identify the corresponding cluster parameter vector $\tilde{\kappa}_{.D_k}$. Within a cluster l, dummies have identical parameter values $\tilde{\kappa}_{.l}$ 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 (4) 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. Technically, d_i can also be a set of continuous variables or binary dummies. However, clustering parameters of continuous variables and binary dummies lacks interpretation in many applications and are therefore often included in w_i.

2.2.2 Parameter clustering over outcome categories

The model in (3) can be written as

$$\eta_{ij} = \alpha_j + \gamma_j' w_i + \sum_{k=1}^{K_d} \kappa_{jk} d_{ik}, \quad j = 1, \ldots, J, \tag{5}$$

where γ_j' is the jth row of γ, and the elements of γ_j and κ_{jk} are alternative-specific parameters. Equivalently, the parameters vary over a number of clusters across choice alternatives.

We combine parameter clustering over explanatory categories in (4) with parameter clustering over outcome categories in a two-way parameter clustering. The two-way cluster representation of (5) is

$$\eta_{ij} = \alpha_j + \tilde{\gamma}'_{C_j} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j,D_k} d_{ik}, \quad j = 1, \ldots, J, \tag{6}$$

where $\gamma_j = \tilde{\gamma}_{C_j}$, $\kappa_{jk} = \tilde{\kappa}_{C_j,D_k}$, and the classification variables $C_j \in \{1, \ldots, L_J\}$ take integer values indicating the cluster for choice category j. For model identification, $C_1 = 1$ and the parameter values in $\tilde{\beta}_1$ equal zero. The coefficients in γ_j can vary over L_J clusters and in κ_{jk} over $L_J \times L_K$ clusters for $j = 2, \ldots, J$, where L_J equals the number of clusters across the J outcome categories and L_K the number of clusters across the $K_d + 1$ explanatory categories.

The parameters in (6) are potentially clustered over the outcome categories except for the intercepts. The odds ratio of alternative j relative to l equals

$$\frac{P(y_i = j|x_i)}{P(y_i = l|x_i)} = \exp(\alpha_j - \alpha_l + x_i'(\beta_j - \beta_l)). \tag{7}$$

In case j and l share a cluster, and $\beta_j = \beta_l$, the odds ratio only includes the intercepts, which take the variation across outcome category frequencies into account.

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 prior 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 prior encourages a parsimonious model. First, we specify a Dirichlet process prior that clusters over explanatory categories. Sec-

ond, we extend the Dirichlet process prior to a two-way clustering prior over explanatory categories and outcome categories.

2.3.1 Dirichlet process prior

A data-driven parameter clustering over explanatory categories is obtained by specifying a prior for the parameter vectors $\kappa_{2:J,k} = (\kappa_{2k}, \ldots, \kappa_{Jk})'$ in (3),

$$\kappa_{2:J,k}|P \sim P, \quad P|\lambda_K, G_\kappa \sim DP(\lambda_K, G_\kappa), \quad G_\kappa = N(0, \sigma_\beta^2 I_{J-1}), \tag{8}$$

where the prior is a random distribution P generated by a Dirichlet process. Conditionally on P, the parameter vectors $\kappa_{2:J,k}$, $k = 1, \ldots, K_d$, are independently and identically distributed. The Dirichlet process $DP(\lambda_K, G_\kappa)$ has a positive scalar concentration parameter λ_K and base distribution G_κ.

The expectation over the Dirichlet process equals the base distribution, and the concentration parameter governs the dispersion around the base distribution. When λ_K is large, the distributions P and G_κ are more similar. Since P is a discrete random distribution, there is a positive probability that different $\kappa_{\cdot k}$s take the exact same value. A cluster of explanatory categories is defined as the explanatory categories k with identical parameter vectors $\kappa_{\cdot k}$.

2.3.2 Stick-breaking representation

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

$$P = \sum_{l=1}^{L_K} p_l \delta(\tilde{\kappa}_{2:J,l}), \qquad \tilde{\kappa}_{2:J,l} \sim G_\kappa, \tag{9}$$

where $L_K \to \infty$ and $\delta(\tilde{\kappa}_{2:J,l})$ denotes a unit-mass measure concentrated at $\tilde{\kappa}_{2:J,l}$. The Dirichlet process is a distribution over independent and identically distributed draws from

the base distribution, with random weights

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

where $V_l \sim \text{Beta}(1, \lambda_K)$. This process is also written as $p = (p_1, \dots, p_{L_K})' \sim \text{stick}(\lambda_K)$. Since $\sum_{l=1}^{L_K} p_l = 1$, 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 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 $\text{E}[V_l] = \frac{1}{1+\lambda_K}$, a small λ_K 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_K, the weights in p are more evenly distributed.

The Dirichlet process prior for $\kappa_{\cdot k}$ in (8) can be equivalently formulated by means of the classification variables $D = (D_1, \dots, D_{K_d})'$ in (4). Using the stick-breaking representation of the Dirichlet process prior in (9), we have

$$\kappa_{2:J,k} = \sum_{l=1}^{L_K} 1[D_k = l]\tilde{\kappa}_{2:J,l}, \;\; D_k \sim \sum_{l=1}^{L_K} p_l \delta(l), \;\; p \sim \text{stick}(\lambda_K), \;\; \tilde{\kappa}_{2:J,l} \sim G_\kappa, \tag{12}$$

where $1[A]$ is an indicator function that equals one if event A occurs and zero otherwise. The model in (3) with the prior in (12) is known as a Dirichlet process mixture model, which in this case clusters over explanatory categories.

2.3.3 Two-way Dirichlet process prior

We use the stick-breaking presentation of the Dirichlet process prior to extend the one-way prior on explanatory categories to a two-way Dirichlet process prior:

$$\kappa_{jk} = \sum_{l=1}^{L_J} \sum_{m=1}^{L_K} 1[C_j = l] 1[D_k = m] \tilde{\kappa}_{lm}, \quad D_k \sim \sum_{m=1}^{L_K} p_m \delta(m), \quad p \sim \text{stick}(\lambda_K)$$
$$\gamma_j = \sum_{l=1}^{L_J} 1[C_j = l] \tilde{\gamma}_l, \quad C_j \sim \sum_{l=1}^{L_J} q_l \delta(l), \quad q \sim \text{stick}(\lambda_J), \quad \tilde{\gamma}_l \sim G_\gamma, \quad \tilde{\kappa}_{lm} \sim G_{\kappa_l},$$
$$k = 1, \ldots, K, \quad j = 2, \ldots, J, \quad l = 1, \ldots, L_J, \quad m = 1, \ldots, L_K, \tag{13}$$

where $q = (q_1, \ldots, q_{L_J})'$. The base distributions are specified as $G_\gamma = N(0, \sigma_\beta^2 I_{K_w})$ and $G_{\kappa_l} = N(0, \sigma_\beta^2)$. Define the $L_J \times (K_w + L_K)$ cluster coefficient matrix $\tilde{\beta} = (\tilde{\gamma}, \tilde{\kappa})$, where $\tilde{\gamma}$ is an $L_J \times K_w$ matrix with rows $\tilde{\gamma}_l'$, and $\tilde{\kappa}$ is an $L_J \times L_K$ matrix with elements $\tilde{\kappa}_{lm}$. The coefficient vector β_j is constructed from $\tilde{\beta}$, C, and D as $\beta_j = (\tilde{\gamma}_{C_j}', \tilde{\kappa}_{C_j,D_1}, \ldots, \tilde{\kappa}_{C_j,D_{K_d}})'$.

The model in (5) with the prior in (13) is a Dirichlet process mixture model, mixing over explanatory categories and outcome categories. This two-way Dirichlet process mixture model has one-way clustering over explanatory or outcome categories as special cases. Instead of sampling the classification variables C_j from a mixing distribution, we set $C = (C_1, \ldots, C_J)' = (1, \ldots, J)'$ and $L_J = J$ for one-way clustering over explanatory categories. This prior can also be applied to a binary logit with $J = 2$. For one-way clustering over outcome categories, we fix $D = (1, \ldots, K_d)'$ and $L_K = K_d$.

3 Bayesian inference

To estimate the parameters, 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. Moreover, since the truncation allows for blocked updates for the cluster weights and classification variables, the sampler mixes relatively well (Ishwaran and James, 2002). Nibbering (2023) shows that the approximation error

from the truncation can be estimated from the output of the Gibbs sampler.

3.1 Prior distributions

The two-way mixture model is defined by a Dirichlet process prior on the parameters β. The prior distribution for the intercepts is specified as

$$\alpha_j|\sigma_\alpha^2 \sim N(0, \sigma_\alpha^2), \quad j = 2, \ldots, J. \tag{14}$$

We let the data determine the number of clusters by treating the concentration parameters as unknown with a prior distribution,

$$\lambda_J|\theta_{J1}, \theta_{J2} \sim \text{Gamma}(\theta_{J1}, \theta_{J2}), \quad \lambda_K|\theta_{K1}, \theta_{K2} \sim \text{Gamma}(\theta_{K1}, \theta_{K2}), \tag{15}$$

where $\text{Gamma}(\theta_{.1}, \theta_{.2})$ denotes a gamma distribution with mean $\theta_{.1}/\theta_{.2}$. The values $(\theta_{.1}, \theta_{.2}) \in \mathcal{R}^+$ directly effect the number of estimated clusters through the concentration parameter, where larger values for $\lambda.$ encourage more distinct values for the coefficients. We set $\theta_{.1}/\theta_{.2}$ equal to the value that matches our prior belief about the mode of the distribution of the number of clusters, and use $\theta_{.2}$ to govern the dispersion around the mean. Details are deferred to Nibbering (2023).

3.2 Posterior distribution

The mixture model in (13) conditions on the choice-alternative classification variable C and the explanatory dummy category classification variables D. This model representation allows for a Markov Chain Monte Carlo sampler that simulates the latent classification variables alongside the intercepts α and the cluster coefficient matrix $\tilde{\beta}$. The simulation scheme consists of the following steps:

1. Initialize λ_J, λ_K, q, p, C, D, α, $\tilde{\beta}$.
2. Sample $\omega|\alpha, \tilde{\beta}, C, D, X$.
3. Sample $\alpha, \tilde{\beta}|C, D, \sigma_\alpha^2, \sigma_{\tilde{\beta}}^2, \omega, y, X$.

4. Sample $C|q, \alpha, \tilde{\beta}, D, y, X$.

5. Sample $q|C, \lambda_J$ and $\lambda_J|q, \theta_{J1}, \theta_{J2}$.

6. Sample $D|p, \alpha, \tilde{\beta}, C, y, X$.

7. Sample $p|D, \lambda_K$ and $\lambda_K|p, \theta_{K1}, \theta_{K2}$.

Steps 2 and 3 follow Polson et al. (2013), who show that the coefficients in logistic models can be sampled conditional on a set of Polya-Gamma latent variables ω. We apply this idea to our multinomial logit model, and sample the two-way cluster coefficient matrix $\tilde{\beta}$ conditional on the classification variables C and D and the latent variables in ω. Steps 5 and 7 are standard Gibbs sampling steps, see for instance Ishwaran and James (2002).

Relative to the sampler for a standard multinomial logit model, Steps 4 and 6 increase the computational costs. Step 4 evaluates the data likelihood for each J and L_J, and Step 6 for each K_d and L_K. These computational costs increase linearly in the dimensions J, K_d, L_J, and L_K. However, since Step 2 and 3 are the most time consuming of the sampling scheme and also required in a standard multinomial logit model, the additional computational costs of parameter clustering is relatively small.

Between iterations of the sampler, the classification values in C and D can switch between clusters. If a label switch occurs during the posterior simulation, statistics such as a cluster-specific posterior mean become uninformative (Frühwirth-Schnatter, 2001; Geweke, 2007; Bauwens et al., 2017). Moreover, since the model parameters have a positive cluster probability for each cluster in each iteration of the sampler, also the number of clusters can change at each sample iteration. When label-switching occurs, label invariant posterior statistics can still be interpreted as long as the sampler visits the entire sample space.

3.3 Parameter clustering in other choice models

To avoid the independence of irrelevant alternatives (IIA) assumption of the multinomial logit model (Hausman and McFadden, 1984), researchers may consider to apply parameter clustering in other choice models.

The multinomial probit model allows the latent variables, similar to η_{ij} in (2), to be correlated across j, and hence avoids the IIA. The likelihood of the multinomial probit model includes an integral over these latents, which is intractable for any choice of covariance matrix (Johndrow et al., 2013). Hence, Bayesian estimation of the model targets the posterior augmented with the latent utilities, and the model parameters are sampled conditional on the normally distributed latent utilities. In combination with parameter clustering, not only the model parameters but also the latent cluster indicators are sampled conditional on a large number ($J \times N$) of latent variables, resulting in slow convergence. In the MNL sampler however, the likelihood can be easily evaluated and the sampling step for the cluster indicators does not depend on the latents but directly on the data y.

Alternatively, the multinomial logit model can be extended with random coefficients of alternative-specific covariates or with random coefficients that are correlated across choice alternatives to avoid the IIA property. These models are known as mixed logit models, and are usually estimated with alternative-specific covariates and repeated observations, as for example in Train (2009), Greene and Hensher (2010), and Fiebig et al. (2010). Although the model is formally identified without repeated observations for each individual, Fiebig et al. (2010) discuss that the estimation of these models is challenging. Hence, provided that there are sufficient repeated observations and alternative-specific covariates available, the mixed logit model can be combined with a two-way Dirichlet process prior on alternative-specific coefficients for individual-specific covariates. Since the sampling steps for the cluster indicators in the MCMC sampler for this model still depend on the data y, as conditional on the random coefficients its likelihood can still be evaluated, it does not suffer from the convergence issues in the sampler for the multinomial probit model with a two-way Dirichlet process prior.

Nibbering (2023) Part A discusses the sampling steps in more detail and explains how the predictive distribution can be computed. Part B examines the practical implications of the proposed parameter clustering methods on simulated data. Part C specifies a mixed logit model with a two-way Dirichlet process prior, discusses its MCMC sampler,

and estimates the model on simulated data. The convergence of the sampler can be assessed using the convergence diagnostics discussed in Part D.

4 Empirical application

This section estimates the relation between household compositions and holiday destinations using survey data from a Dutch market research company. The company is interested in how preferences for destinations differ across household types. Nibbering (2023) includes additional details on the data and the results.

4.1 Data

The data set consists of 14,661 reported holidays undertaken in 2015 by 6512 Dutch respondents and their individual characteristics. 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 holidays with a foreign destination of more than seven days. This data set contains 4907 holiday destinations of 3334 households. Jointly analyzing the decision process for the 1881 domestic holidays and the 4907 foreign holidays requires a baseline inflated choice model, which is outside the scope of this paper.

The survey does not include information on different holiday locations within the same destination. Hence, travelers who stayed for more than one week in one destination are counted as one holiday, also if they stayed in different locations within the same destination. Holiday stays that lasted more than two weeks are counted as one long stay. As different locations within one destination could be seen as two trips, information on the specific holiday locations may prevent a potential bias in our findings. The minimum, median, mean, and maximum holiday spell across all holiday destinations equal respectively 2, 6, 8.818, and 89 days in the full data sample, and 8, 13, 14.835, and 89 days in the sample used for estimation.

The respondents 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. We delete the ten categories which are never chosen, as the corresponding coefficients are not identified. Since a marketing manager is less interested in low volume destinations, and categories with only a few observations substantially slow down the convergence of the sampler, we group categories to ensure a minimum of ten observations. At the cost of an increase in autocorrelation in the sampler, the minimum number of observations per category can be set to a smaller number. 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.

Note that the holiday destinations may violate the IIA property of the multinomial logit model. This property imposes that the relative probability of a household choosing between two holidays is independent of any additional holiday alternatives in the choice set after conditioning on covariates. For instance, a household that is looking for a beach resort holiday may consider several sunny holiday destinations as substitutes. The relative likelihood of choosing sunny location j over location l may depend on the presence of another sunny location k, if the observed characteristics of the households in the model do not explain the preference for sunny holidays. In such a case, the parameter estimates may be biased. Hence, it is important to include a large set of individual-level controls in the model, that captures a substantial amount of individual heterogeneity in holiday preferences. In case there is little individual-specific information available, the IIA assumption can be relaxed with random coefficients in the presence of sufficient repeated observations and alternative-specific covariates, as discussed in Section 3.3.

We cluster parameters across the categories of one explanatory variable. Respondents select their household composition out of 11 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 in the tenth category and

older than 35 in the eleventh.

Figure 2 shows the frequency counts for the dummy categories. We include dummy variables for the first ten categories in the model and treat the eleventh as baseline category. In applications in which the computational cost of the Bayesian sampler is a concern, small categories may be merged together a priori to reduce the number of parameters to be estimated, and potentially reducing the autocorrelation in the sampler so that less iterations are required. Using an informative prior distribution on the coefficients β deals with potential multicollinearity with many dummy categories and many additional controls.

We do not cluster parameters across a set of nine control variables. We include one continuous variable, which is log income of the household, and eight 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. This application does not include choice-specific variables in the model. However, the intercepts take the variation across the holiday destinations into account.

Table 1 shows the smallest number of holiday destinations that cover 50%, 80%, 90%, or 100% percent of the observed holiday destination choices, for households with different characteristics. These descriptive statistics suggest that especially families with kids choose from a limited set of holiday destinations.

Since we have 49 outcome categories, an intercept, nine control variables, and ten categorical dummies, the total number of parameters to be estimated is $48 \times 20 = 960$. We randomly select 80% of the holidays for parameter estimation and use the remaining holidays for out-of-sample analysis. Hence, each category has on average 80 observations, and the smallest category only 10 observations, to estimate 20 parameters.

4.2 Modelling choices

We estimate four models; a two-way mixture model, one-way mixture models that either cluster over outcome or explanatory categories, and a standard multinomial logit (MNL) model. For all models, the prior specification for the parameters α and β follows (13)

Table 1: Descriptive statistics holiday destination choices by household categories

minimum #	social class A and B1			social class B2, C and D			total
destinations for	no kids	teens	kids	no kids	teens	kids	sample
50% observations	8	5	4	4	5	4	5
80% observations	20	14	9	10	14	7	13
90% observations	28	23	16	16	23	11	24
100% observations	44	49	44	21	49	31	49
total observations	250	2134	839	58	1364	262	4907

This table shows the smallest number of holiday destinations that cover 50%, 80%, 90%, or 100% percent of the observed holiday destination choices, for households with different characteristics. The first three columns correspond to households with either no kids ('single <35', 'hh>1, head<35'), teens ('single =>35', 'hh>1, head>=35', 'kids 13-17'), or kids ('kids 0-5, 13-17', 'kids 6-12', 'kids 0-5', 'kids 0-5, 6-12', 'kids 6-12, 13-17', 'kids 0-5, 6-12, 13-17') from social class A and B1, the next three columns to the same three categories from social class B2, C, or D, and the final column to the full sample. The final row shows the number of observations in each category.

and (14), with $\sigma_\alpha^2 = \sigma_\beta^2 = 1$. This allows for a wide range of plausible values for model parameters within a multinomial choice model.

Our prior belief about the mode of the number of clusters over holiday destinations is fifteen and over household compositions five. The prior distributions that match these beliefs are $\lambda_J \sim \text{Gamma}(7.15 \times 20, 20)$ and $\lambda_K \sim \text{Gamma}(3.47 \times 1, 1)$, with a variance of nine and three, respectively. The truncation level of the number of potential choice category clusters is set equal to $L_J = 25$. We do not truncate the number of potential dummy category clusters: $L_K = K_d + 1 = 11$.

The performance of the models is evaluated by the log-score and the hit-rate. The log-score equals

$$\text{LS} = \sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \ln(\hat{Pr}(y_i = j)), \tag{16}$$

where $\hat{Pr}(y_i = j) = \frac{1}{S} \sum_{s=1}^{S} I[y_i^{(s)} = j]$, and S is the number of draws from the predictive distribution defined in Nibbering (2023). The hit-rate is defined as

$$\text{HR} = \frac{1}{N} \sum_{i=1}^{N} I[\hat{y}_i = y_i], \tag{17}$$

where $\hat{y}_i$ is the mode of the draws from (41). For both the log-score and the hit-rate large

values are preferred. We test the difference of the log-scores between models with the Giacomini and White (2006) test and the difference of the hit-rates with a normal test. Since only the log-score is a strictly proper scoring rule (Gneiting and Raftery, 2007), we consider this as the preferred metric for model evaluation.

Posterior results are based on 1,000,000 iterations of the Gibbs sampler, from which the first 500,000 are discarded, and we use a thinning value of 50.

4.3 Results

Table 2 shows the in- and out-of-sample performance for predicting holiday destinations. The symbol (*) indicates that only the out-of-sample log-score of the two-way mixture model significantly improves upon the standard MNL. This improvement costs an 8.9% increase in computation time. The out-of-sample log-scores of the other mixture models also improve relative to the standard MNL, but the difference is not significant. The standard MNL has the largest in-sample log-score, which can be explained by the high degrees of freedom. The differences between the models in the hit-rates are not significant. All models improve substantially on the naive forecast, which is constructed as the observed in-sample frequency of the choice alternatives. Estimating the standard MNL as a set of binary logit models gives similar results.

The two-way mixture model improves the log-score upon the benchmark models by estimating less unique model parameter values. Figure 3 shows that the mode of the distribution over the number of clusters for holiday destinations equals seven, and for household types four, in the two-way mixture model. These numbers of clusters would result in 126 unique parameter values. This is a substantial reduction relative to 960 unique parameter values in the standard multinomial logit model, 564 in the one-way mixture model over holiday destinations, and 624 in the one-way mixture model over household types.

The decrease in the estimated number of unique parameters, decreases the posterior uncertainty. Figure 4 shows the posterior distributions of three parameters: the intercept, the coefficient of income, and the coefficient of the dummy for the household category

Table 2: Model evaluation

		clustering			standard	
sample	metric	two-way	holiday	household	MNL	naive
in	log-score	-2.951	-2.967	-2.913	-2.891	-3.042
in	hit-rate	0.195	0.193	0.192	0.192	0.148
out	log-score	-2.977*	-2.986	-2.992	-3.000	-3.030
out	hit-rate	0.202	0.204	0.204	0.205	0.177
	time	1.089	1.192	1.039	1.000	

This table shows in-sample and out-of-sample performance for predicting holiday destinations measured by log-scores and hit-rates as defined in (16) and (17), respectively. The performance of the two-way mixture model is compared to a one-way mixture model that clusters holiday destinations, a one-way mixture model that clusters household categories, a standard MNL 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 symbol (*) indicates that the method indicated by the column label performs significantly better than the standard MNL model, on a significance level of 5%. The final row shows the computation time of the models relative to the standard MNL model.

with kids between 0 and 5 years old, for the holiday destination the United Kingdom. The posterior standard deviations in the mixture models are smaller than in the standard multinomial logit model.

We now zoom into the results of the two-way mixture model. Figure 5 shows the posterior probabilities that household categories cluster together. The pairwise probabilities distinguish three segments of households: younger than 35 without kids, 35 or older without kids or with teenagers, and households with kids younger than 13. The pairwise cluster probabilities vary between 50.9% and 96.2% within these segments, and between 0.0% and 12.8% between the segments. Household categories with cluster probabilities close to one have a similar preference ranking across holiday destinations, conditional on the control variables. For instance, singles younger than 35 and households younger than 35 –with a pairwise cluster probability of 96.2%– have almost identical choice probabilities.

The cluster probabilities of the household categories may be used to design effective marketing strategies. Since it is costly to target each individual household type with a different marketing strategy, the cost-effectiveness may be increased by only differentiating between a small number of household segments that differ most from each other in

terms of holiday preferences. For instance, personalizing holiday flyers for each household type is costly, but sending a different flyer to each household segment may be feasible. Figure 5 suggests that targeting three segments of households –younger than 35 without kids, 35 or older without kids or with teenagers, and households with kids younger than 13– may be more cost-effective than targeting eleven different household categories.

Moreover, the amount of information that can be collected from future holiday makers may be limited as, for instance, potential customers are only willing to answer a small number of quick questions on the landing page of a travel agency website. Figure 5 shows that the simple questions (1) "Do you have kids younger than thirteen?" and if not, (2) "Is the head of your household younger than 35?" are able to distinguish three segments of households with similar holiday preferences.

Figure 6 shows the pairwise posterior probabilities that holiday destinations cluster together. Each destination has multiple pairwise cluster probabilities larger than 5%, which shows that the model applies shrinkage to each choice alternative. In case a pairwise cluster probability equals one, the corresponding odds ratio is only based on the intercepts, and hence the same for all households. We find eleven unique cluster probabilities larger than 95%, which correspond to alternatives with very similar coefficients: for instance, Turkey and Greece, Southern-America and Mexico, and Netherlands Antilles and United States. The model does not (strongly) shrink the coefficients toward each other for the holiday destinations corresponding to the 36.7% of pairwise probabilities that are smaller than 5%.

Figure 7 shows the odds ratios, as defined in (7), for Turkey, Greece, Netherlands Antilles, and the United States. The odds ratios are calculated for singles younger than 35, households younger than 35, households with kids in the age group 0-5, and with kids between 0-5 and 6-12, with the following values for the control variables: mean log income, not retired or student, no fixed or moving holiday accommodation, and social class A. The odds ratios are similar between households within the segments identified in Figure 5, and different between households across these segments.

Travel agencies that offer a large number of holiday destinations can use the cluster

probabilities in Figure 6 and the odds ratios in Figure 7 to focus marketing efforts on a small set of destinations. For instance, the cluster probability between Turkey and Greece equals 97.9%, and between the Netherlands Antilles and the United States 92.8%. The remaining cluster probabilities between these four destinations are smaller than 2%. Since a pair of destinations with a large cluster probability has similar coefficients, household characteristics have a small impact on the corresponding odds ratio. Figure 7 shows indeed that the odds-ratio of Turkey and Greece, and the Netherlands Antilles and the United States, are very similar across household categories, and respectively in favour of Greece and the the United States. To promote for each household the destination with the highest choice probability among the four destinations, all marketing efforts can be allocated to Greece and the United States.

A standard multinomial logit model does not provide cluster probabilities, which increases the complexity of designing a focused and effective marketing strategy. It is challenging to extract segments of households that have similar preferences across clusters of holiday destinations from a large number of estimated coefficients. Moreover, the estimated coefficients itself also exhibit more variation, as the standard multinomial logit does not shrink households and holiday destinations towards each other. Figure 8 illustrates this by means of the odds ratios from the standard multinomial logit. These odds ratios vary substantially among all household categories and holiday destinations, in contrast to the odds ratios of the two-way mixture model in Figure 7.

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.

In an empirical application to 49 holiday destinations and 11 household categories, we find that the two-way mixture model substantially reduces the number of unique parameter values by clustering over holiday destinations and household categories. The resulting decrease in parameter uncertainty significantly improves the out-of-sample log-score relative to the standard multinomial logit model. The estimated model parameters and cluster probabilities identify three segments of households that have different preferences across the holiday destinations, and sets of holiday destinations for which all households have similar preferences.

The estimates in the two-way mixture model can be used to design an effective marketing strategy. For instance, marketing efforts can be differentiated across the three segments of households with different holiday preferences, instead of all eleven household types. Advertising the set of holiday destinations for which all households have similar preferences, do not require targeting of specific households.

The results may also guide data collection on potential future customers. On the one hand, the model suggests that only collecting the information that distinguishes household segments, instead of all individual household types, may be sufficient for an effective marketing strategy. On the other hand, the model could possibly be extended with a set of explanatory variables that can differentiate between holidays with high cluster probabilities. The model may also identify holiday offerings that can possibly be removed from the assortment. For each pair of holidays with a high cluster probability, one holiday is dominated in terms of choice probability for all households, and the cost effectiveness of taking it out of the assortment may be worth investigating.

References

Agresti, A. *Categorical data analysis*, volume 482. John Wiley & Sons, 2003.

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.

Becg, C. B. and Gray, R. Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71(1):11–18, 1984.

Bhadra, A., Datta, J., Polson, N. G., Willard, B., et al. Lasso meets horseshoe: a survey. *Statistical Science*, 34(3):405–427, 2019.

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

Brooks, S. P. and Gelman, A. General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics*, 7(4):434–455, 1998.

Burgette, L. F., Reiter, J. P., et al. Multiple-shrinkage multinomial probit models with applications to simulating geographies in public use data. *Bayesian Analysis*, 8(2): 453–478, 2013.

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

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.

Fiebig, D. G., Keane, M. P., Louviere, J., and Wasi, N. The generalized multinomial logit model: accounting for scale and coefficient heterogeneity. *Marketing science*, 29 (3):393–421, 2010.

Frühwirth-Schnatter, S. Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96 (453):194–209, 2001.

Gelman, A. and Rubin, D. B. Inference from iterative simulation using multiple sequences. *Statistical science*, 7(4):457–472, 1992.

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.

Geweke, J. Interpretation and inference in mixture models: Simple MCMC works. *Computational Statistics & Data Analysis*, 51(7):3529–3550, 2007.

Geweke, J., Keane, M., and Runkle, D. Alternative computational approaches to inference in the multinomial probit model. *The review of economics and statistics*, pages 609–632, 1994.

Geweke, J., Gowrisankaran, G., and Town, R. J. Bayesian inference for hospital quality in a selection model. *Econometrica*, 71(4):1215–1238, 2003.

Giacomini, R. and White, H. Tests of conditional predictive ability. *Econometrica*, 74 (6):1545–1578, 2006.

Gneiting, T. and Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *Journal of the American statistical Association*, 102(477):359–378, 2007.

Greenberg, E. *Introduction to Bayesian econometrics*. Cambridge University Press, 2012.

Greene, W. H. and Hensher, D. A. Does scale heterogeneity across individuals matter? An empirical assessment of alternative logit models. *Transportation*, 37(3):413–428, 2010.

Hausman, J. and McFadden, D. Specification tests for the multinomial logit model. *Econometrica: Journal of the econometric society*, pages 1219–1240, 1984.

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

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.

Johndrow, J., Dunson, D. B., and Lum, K. Diagonal orthant multinomial probit models. In *AISTATS*, pages 29–38, 2013.

MacLehose, R. F. and Dunson, D. B. Bayesian semiparametric multiple shrinkage. *Biometrics*, 66(2):455–462, 2010.

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

Nibbering, D. Online supplementary appendix to "A high-dimensional multinomial logit model". *Journal of Applied Econometrics*, 2023.

Pauger, D., Wagner, H., et al. Bayesian effect fusion for categorical predictors. *Bayesian Analysis*, 14(2):341–369, 2019.

Polson, N. G., Scott, J. G., and Windle, J. Bayesian inference for logistic models using Pólya–G amma latent variables. *Journal of the American statistical Association*, 108 (504):1339–1349, 2013.

Sethuraman, J. A constructive definition of Dirichlet priors. *Statistica Sinica*, 4:639–650, 1994.

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

Train, K. E. *Discrete choice methods with simulation*. Cambridge university press, 2009.

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

Vincent, M. and Hansen, N. R. Sparse group lasso and high dimensional multinomial classification. *Computational Statistics & Data Analysis*, 71:771–786, 2014.

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

A Notation

Multinomial logit model	
y_i	integer indicating choice of individual i
x_i	$K \times 1$ vector with characteristics of individual i
η_i	$J \times 1$ vector with the linear predictor functions of individual i
w_i	$K_w \times 1$ vector with variables for which we do not cluster the parameters
d_i	$K_d \times 1$ vector with dummy variables for which we cluster the parameters
α	$J \times 1$ vector of intercepts
β	$J \times K$ coefficient matrix of x_i
γ	$J \times K_w$ coefficient matrix of w_i
κ	$J \times K_d$ coefficient matrix of d_i
Dimensions	
J	number of choice alternatives
N	number of individuals
K	number of explanatory variables
K_w	number of explanatory variables for which we do not cluster the parameters
K_d	number of dummy variables for which we cluster the parameters
Parameter clustering	
L_K	number of clusters for the explanatory categories
L_J	number of clusters for the outcome categories
$\tilde{\beta}$	$L_J \times (K_w + L_K)$ cluster coefficient matrix of x_i
$\tilde{\gamma}$	$L_J \times K_w$ cluster coefficient matrix of w_i
$\tilde{\kappa}$	$L_J \times L_K$ cluster coefficient matrix of d_i
D_k	classification variable of explanatory category k
C_j	classification variable of outcome category j
Dirichlet process mixture	
λ_K	concentration parameter explanatory categories
λ_J	concentration parameter outcome categories
G_κ	base distribution κ
G_γ	base distribution γ
p_l	probability of explanatory category cluster l
q_l	probability of outcome category cluster l

B Multiple explanatory categorical variables

This appendix extends the clustering methods for explanatory categories to multiple categorical explanatory variables in $x_i = (w_i', d_{i1}', \ldots, d_{iH}')'$. The c_h dummy variables in $d_{ih} = (d_{ih1}, \ldots, d_{ihc_h})'$ correspond to the first c_h of the $c_h + 1$ categories in a categorical explanatory variable. We rewrite the model in (3) to

$$\eta_i = \alpha + \gamma w_i + \sum_{h=1}^{H} \kappa_{.h} d_{ih} + \varepsilon_i = \alpha + \gamma w_i + \sum_{h=1}^{H} \sum_{k=1}^{c_h} \kappa_{.hk} d_{ihk}, \tag{18}$$

where $\beta = (\gamma, \kappa_{.1}, \ldots, \kappa_{.H})$ with $\kappa_{.h} = (\kappa_{.h1}, \ldots, \kappa_{.hc_h})$ for $h = 1, \ldots, H$. The parameter values in $\kappa_{.hk} = (\kappa_{1hk}, \ldots, \kappa_{Jhk})'$ for $k = 1, \ldots, c_h$ correspond to the dummy variable d_{ihk}. We cluster the parameters $\kappa_{.hk}$ across the categories within the categorical explanatory variable d_{ih}, for each $h = 1, \ldots, H$ in (18).

The classification variables $D_{hk} \in \{1, \ldots, L_{c_h}\}$ for category k in explanatory categorical variable h, identify the cluster parameter vector $\kappa_{.hk} = \tilde{\kappa}_{.hD_{hk}}$. Using the stick-breaking representation of the Dirichlet process prior, we have

$$\kappa_{.hk} = \sum_{l=1}^{L_{c_h}} 1[D_{hk} = l] \tilde{\kappa}_{.hl}, \;\; D_{hk} \sim \sum_{l=1}^{L_{c_h}} p_{hl} \delta(l), \;\; p_h \sim \text{stick}(\lambda_{c_h}), \;\; \tilde{\kappa}_{.hl} \sim G_{\kappa_{.h}}, \tag{19}$$

with L_{c_h} clusters, concentration parameter λ_{c_h}, and base distribution $G_{\kappa_{.h}}$.

The two-way mixture model with multiple categorical explanatory variables is

$$\eta_{ij} = \alpha_j + \gamma_j' w_i + \sum_{h=1}^{H} \sum_{k=1}^{c_h} \kappa_{jhk} d_{ihk}, \tag{20}$$

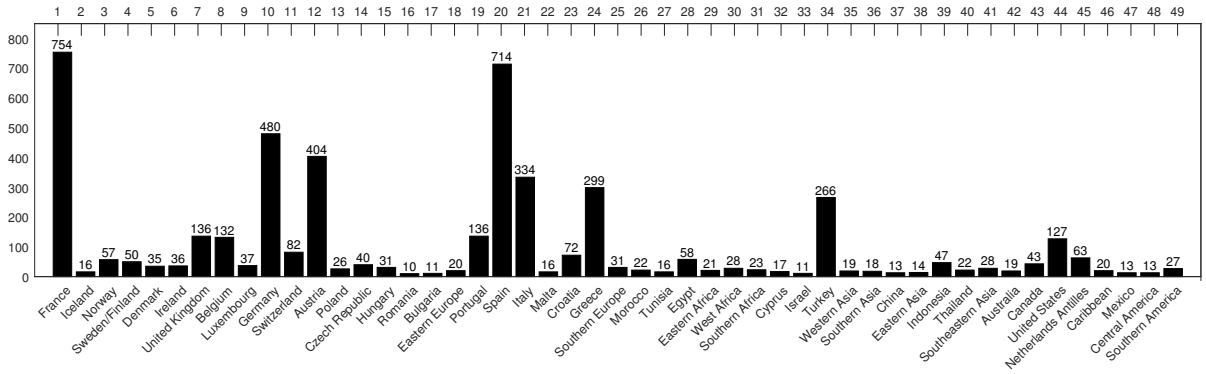
$$\gamma_j = \sum_{l=1}^{L_J} 1[C_j = l] \tilde{\gamma}_l, \quad \kappa_{jhk} = \sum_{l=1}^{L_J} \sum_{m=1}^{L_{c_h}} 1[C_j = l] 1[D_{hk} = m] \tilde{\kappa}_{lhm}, \tag{21}$$

$$C_j \sim \sum_{l=1}^{L_J} q_l \delta(l), \quad q \sim \text{stick}(\lambda_J), \quad D_{hk} \sim \sum_{m=1}^{L_{c_h}} p_{hm} \delta(m), \quad p_h \sim \text{stick}(\lambda_{c_h}), \tag{22}$$

$$\tilde{\gamma}_l \sim G_\gamma, \quad \tilde{\kappa}_{lhm} \sim G_{\kappa_h}, \quad l = 1, \ldots, L_J, \quad m = 1, \ldots, L_{c_h}. \tag{23}$$

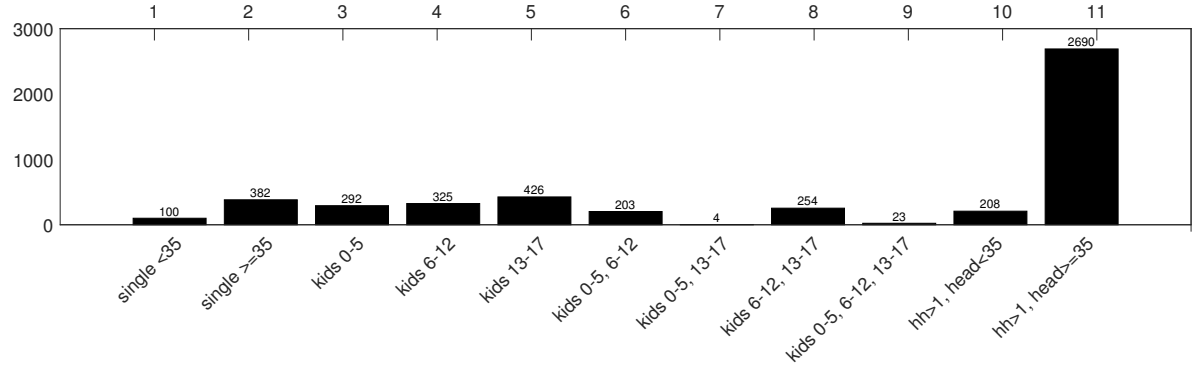
When the categories in only one variable are clustered, $H = 1$, $c_1 = K_d$, $\lambda_{c_1} = \lambda_K$.

Figure 1: Frequency counts choice categories



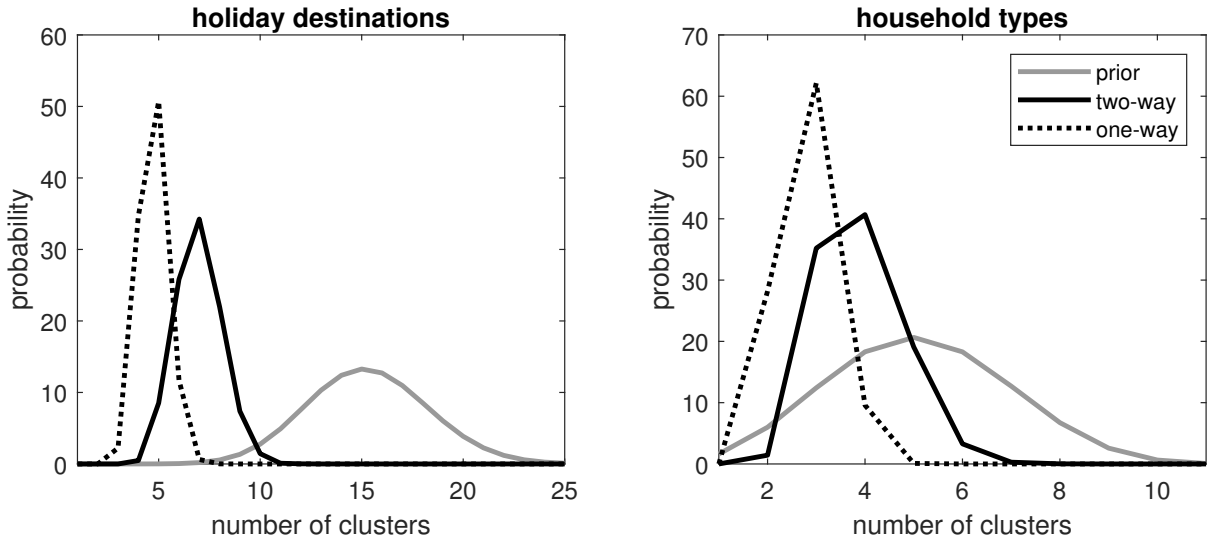
This figure shows the frequency counts for the categorical dependent variable. The categories represent destinations of foreign holidays of more than seven days of Dutch households.

Figure 2: Frequency counts household categories



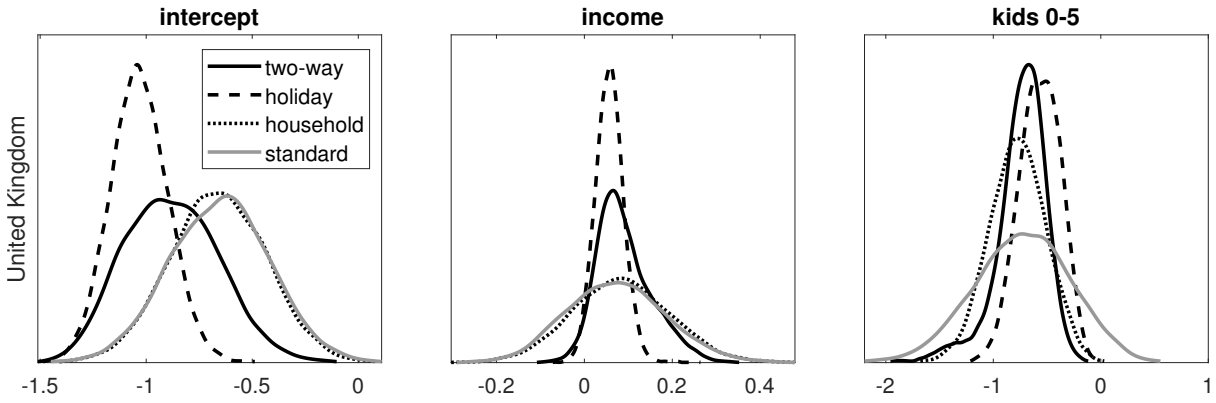
This figure shows the frequency counts for the categorical explanatory variable. The categories represent the household compositions of the survey respondents for each holiday.

Figure 3: Distribution of the number of unique parameter values



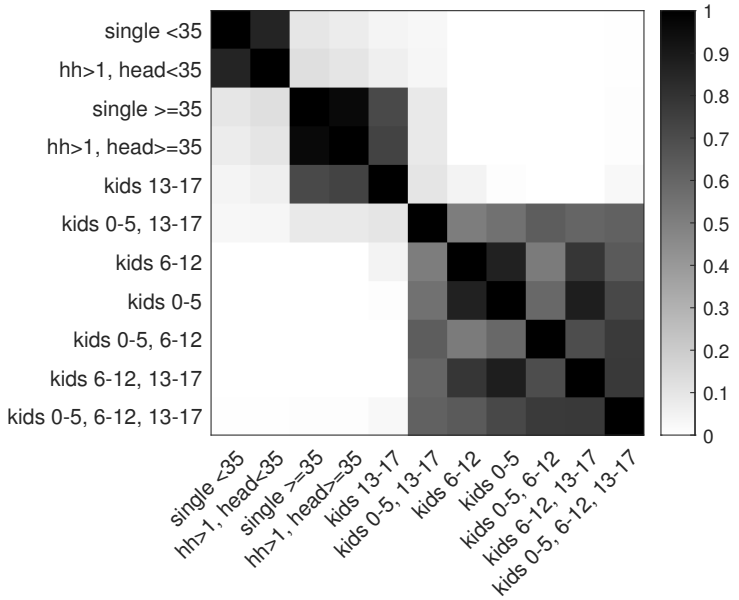
This figure shows the prior distribution (solid gray line), posterior distribution in the two-way mixture model (solid black line), and posterior distribution in the one-way mixture models (dashed black line), over the number of unique parameter values over holiday destinations (left panel) and household compositions (right panel).

Figure 4: Posterior parameter distributions



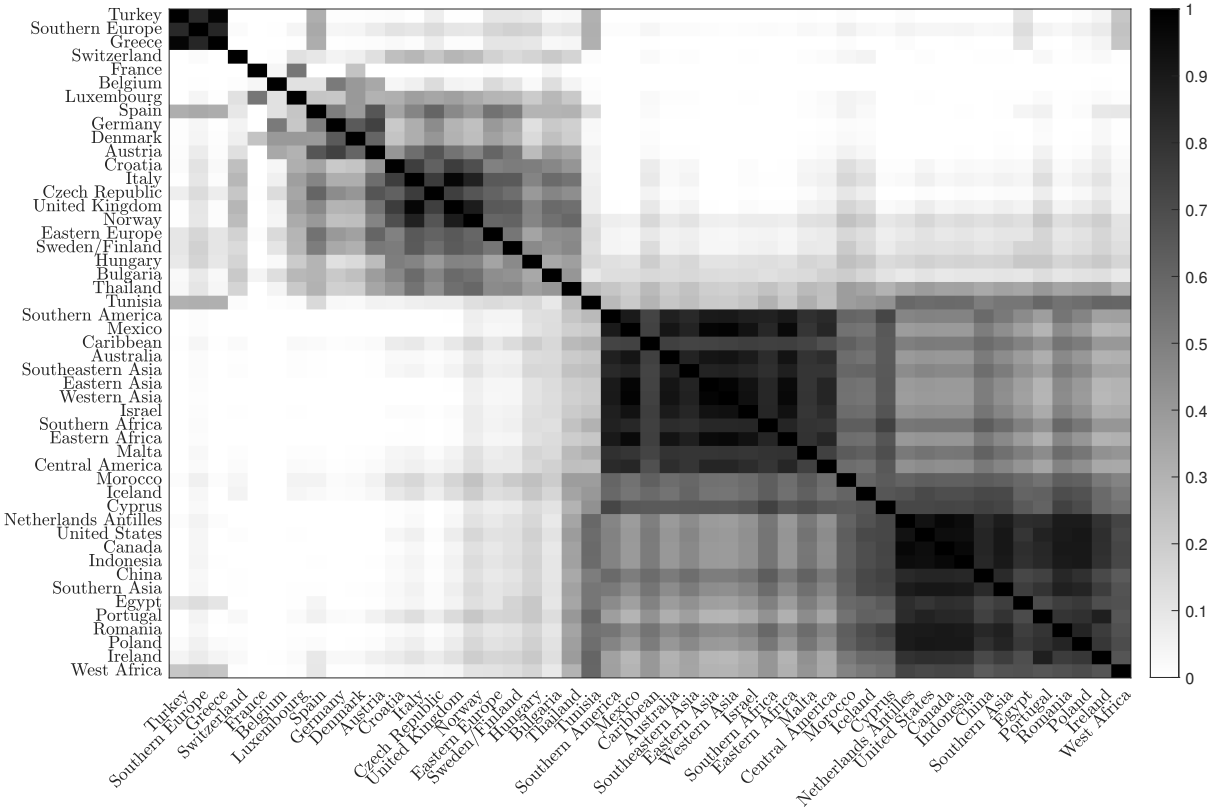
This figure shows the posterior parameter distributions of the intercept (left panel), the coefficient of the control variable for income (middle panel), and the coefficient of the dummy variable for the household category with kids between 0 and 5 years old (right panel), for the holiday destination the United Kingdom. The posterior parameter distributions are estimated in the two-way mixture model (solid black line), the one-way mixture model that clusters holiday destinations (dashed black line), the one-way mixture model that clusters household categories (dotted black line), and the standard multinomial logit model (solid gray line).

Figure 5: Pairwise cluster probabilities for the household categories



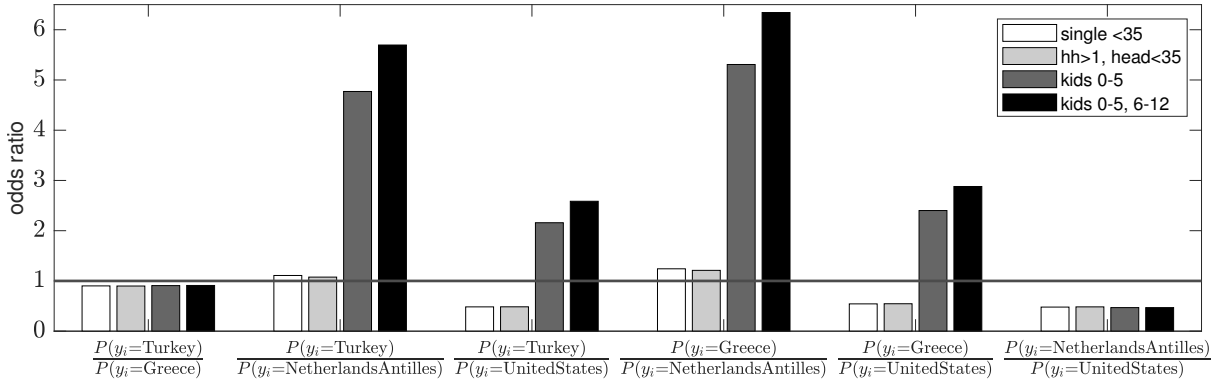
This figure shows the posterior probabilities that the household category at a specific row is in the same cluster as the household category at a specific column in the two-way mixture model. The posterior probabilities range from zero (white) to one (black).

Figure 6: Pairwise cluster probabilities for the holiday destinations



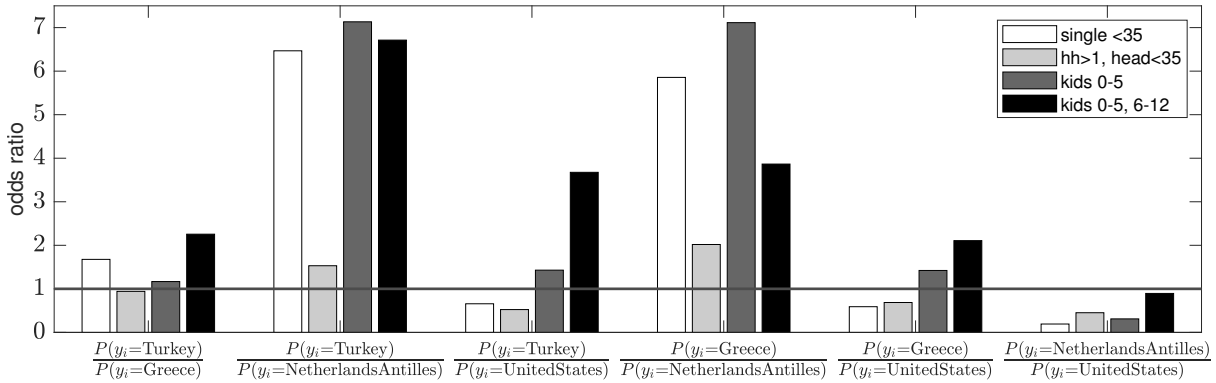
This figure shows the posterior probabilities that the holiday destination at a specific row is in the same cluster as the holiday destination at a specific column in the two-way mixture model. The posterior probabilities range from zero (white) to one (black).

Figure 7: Posterior odds ratios in the two-way mixture model



This figure shows the posterior odds ratios for all combinations of the holiday destinations Turkey, Greece, Netherlands Antilles and United States, for four different household categories. The control variables are set to mean log income, not retired or student, no fixed or moving holiday accommodation, and social class A.

Figure 8: Posterior odds ratios in the standard multinomial logit model



This figure shows the posterior odds ratios for all combinations of the holiday destinations Turkey, Greece, Netherlands Antilles and United States, for four different household categories. The control variables are set to mean log income, not retired or student, no fixed or moving holiday accommodation, and social class A.

Online Supplementary Appendix to 'A high-dimensional multinomial logit model'

This Online Supplementary Appendix has four parts:

Part A: Additional details on the Bayesian inference method.

Part B: Numerical experiments.

Part C: Mixed logit model.

Part D: Additional details on the empirical application.

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 $L = L_K$ for outcome and $L = L_J$ explanatory categories. When the higher order probabilities $\{p_l\}_{l=L}^{\infty}$ in (9) are small enough, the approximation error is negligible. Ishwaran and Zarepour (2000) derive the moments of $\sum_{l=L}^{\infty} p_l$,

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

which are the mean and the variance of the tail probability, respectively, and $\lambda = \lambda_J$ corresponding to J outcome categories or $\lambda = \lambda_K$ corresponding to K_d explanatory categories. These statistics can be used to test whether a truncation level results in a small enough approximation error for a particular λ.

A.2 Concentration parameter

Suppose we have a prior belief about the number of clusters L^*. Van den Hauwe (2015) proposes to set $\lambda = \lambda_J$ corresponding to J outcome categories, or $\lambda = \lambda_K$ corresponding to K_d explanatory categories, to a value that sets $\mathrm{mode}[L^*] = m^*$,

$$\lambda = \frac{1}{2}\left(\exp(-\delta c(m^*+1)) + \exp(-\delta c(m^*))\right), \tag{24}$$

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

Choosing λ as in (24) controls the prior mode of the number of clusters. Conley et al. (2008) show that a fixed concentration parameter may results in a tight prior on the number of clusters. By putting a prior on the concentration parameter, we can also govern the prior variance around the number of clusters.

We specify a prior distribution $f(\lambda)$ with prior mean equal to the value in (24).

To check the dispersion around the prior mode of L^*, we evaluate the marginal prior probability density function with Monte Carlo integration,

$$f(L^*) = \int f(L^*|\lambda) f(\lambda) d\lambda, \tag{25}$$

where $f(L^*|\lambda)$ is the probability function derived by Antoniak (1974),

$$f(L^*|\lambda) = Pr[L^* = j|\lambda] = c(j, J) J! \lambda^j \frac{\Gamma(\lambda)}{\Gamma(\lambda + J)}, \tag{26}$$

for which Escobar and West (1995) discuss how the factors $c(j, J)$ are calculated.

A.3 Posterior simulation

This appendix provides details on the sampling steps for parameter estimation in the two-way Dirichlet process mixture, as discussed in Section 3.2.

A.3.1 Initialization of the sampler

The initial draw for the concentration parameters is $\lambda_J|\theta_{J1}, \theta_{J2} \sim \text{Gamma}(\theta_{J1}, \theta_{J2})$ and $\lambda_K|\theta_{K1}, \theta_{K2} \sim \text{Gamma}(\theta_{K1}, \theta_{K2})$, and for the latent variables $q|\lambda_J \sim \text{stick}(\lambda_J)$, $C_j|q \sim \sum_{l=1}^{L_J} q_l \delta(l)$, $p|\lambda_K \sim \text{stick}(\lambda_K)$, and $D_k|p \sim \sum_{l=1}^{L_K} p_l \delta(l)$. We initialize $\alpha_j = \log\left(\frac{\sum I[y_i=j]}{\sum I[y_i=1]}\right)$ for $j = 2, \ldots, J$ and set the elements of $\tilde{\beta}$ to zero. Given $D = (D_1, \ldots, D_{K_d})$, define the K^*-dimensional vector $x_i^* = (w_i', d_i^{*\prime})'$, with $d_i^* = (\sum_{k=1}^{K_d} I[D_k = D_1^*] d_{ik}, \ldots, \sum_{k=1}^{K_d} I[D_k = D_{m_d}^*] d_{ik})'$ where $D^* = \{D_1^*, \ldots, D_{m_d}^*\}$ denote the current m_d unique values of D.

A.3.2 Sample the latent variables ω

To sample the coefficients α and β, we rewrite the multinomial logit model to $J-1$ binary logistics regressions,

$$P(y_i^{(j)} = j|x_i) = \frac{\exp(\eta_{ij})}{1 + \exp(\eta_{ij})}, \tag{27}$$

where $j > 1$, and $y_i^{(j)}$ equals one if $y_i = j$ and zero if $y_i = 1$, for all i for which $y_i \in \{1, j\}$. In each binary logit, the coefficients α_j and β_j can be sampled conditional on Polya-Gamma latent variables (Polson et al., 2013).

These latent variables $\omega_i^{(j)}$ are sampled as

$$\omega_i^{(j)} | \alpha_j, \tilde{\beta}, C, D, x_i \sim \text{PG}(1, \eta_{ij}), \tag{28}$$

for all i for which $y_i \in \{1, j\}$, and for $j = 2, \ldots, J$. Define $\omega^{(j)}$ as the $N^{(j)}$-dimensional vector, with elements $\omega_i^{(j)}$ corresponding to i for which $y_i \in \{1, j\}$, and $N^{(j)} = \sum_{i=1}^{N} 1[y_i = 1 \text{ or } y_i = j]$. The N_l-dimensional block diagonal matrix Ω_l stacks the $\omega^{(j)}$ with j for which $C_j = l$ on the diagonal, where $N_l = \sum_{j=2}^{J} 1[C_j = l] \sum_{i=1}^{N} 1[y_i = 1 \text{ or } y_i = j]$. The set of all latent variables is denoted as $\omega = \{\omega^{(j)}\}_{j=2}^{J}$.

A.3.3 Sample the model parameters α and $\tilde{\beta}$

First, the coefficients are sampled per nonempty outcome cluster l. Define $\zeta^{(j)}$ as an $N^{(j)}$-dimensional vector, with elements $y_i^{(j)} - 0.5$ corresponding to i for which $y_i \in \{1, j\}$. The N_l-dimensional vector ζ_l stacks the $\zeta^{(j)}$ with j for which $C_j = l$. The rows of the $N_l \times K^*$ regressor matrix X_l contain the $x_i^{*\prime}$ corresponding to the rows in ζ_l. The $N_l \times (\sum_{j=2}^{J} 1[C_j = l] + K^*)$ matrix $Z_l = (A_l, X_l)$ concatenates the regressor matrix X_l to the $N_l \times (\sum_{j=2}^{J} 1[C_j = l])$ matrix A_l with intercepts corresponding to the categories $j > 1$ in cluster l: its rows contain zeros except for the column corresponding to $y_i = j$. For $l = 1$, we have $Z_1 = A_1$. For all nonempty outcome clusters l,

$$(\tilde{\alpha}_l', \tilde{\beta}_l')' | C, D, \sigma_\alpha^2, \sigma_\beta^2, \omega, y, X \sim N(b_l, B_l), \tag{29}$$

with $b_l = B_l Z_l' \zeta_l$, $B_l = (Z_l' \Omega_l Z_l + V_b)^{-1}$, where V_b is a diagonal matrix with σ_α^{-2} as the first $\sum_{j=2}^{J} 1[C_j = l]$ elements and σ_β^{-2} as the final K^* elements on the diagonal. The vector $\tilde{\alpha}_l$ contains the intercepts α_j for all j with $C_j = l$, and $\beta_j = \tilde{\beta}_l$ for all j with $C_j = l$.

Second, the cluster coefficients for the empty outcome clusters are sampled from the

base distribution:

$$\tilde{\beta}_l' | C, D, \sigma_\beta^2 \sim N(0, \sigma_\beta^2 I), \tag{30}$$

for all empty outcome clusters l.

Third, the coefficients corresponding to the empty explanatory clusters are also sampled from the base distribution

$$\tilde{\kappa}_{lk} | C, D, \sigma_\beta^2 \sim N(0, \sigma_\beta^2), \tag{31}$$

for all outcome clusters l and all empty explanatory clusters k.

A.3.4 Sample the classification variables C

Sample the classification variables of the outcome categories as

$$C_j | q, \alpha, \tilde{\beta}, D, y, X \sim \sum_{l=1}^{L_J} \pi_{lj} \delta_l, \tag{32}$$

for $j = 2, \ldots, J$. The conditional cluster probability π_{lj} is a function of the unconditional cluster probability q_l and the data likelihood:

$$(\pi_{1j}, \ldots, \pi_{L_J,j}) \propto \left(q_1 f(\ddot{\beta}_{j1}), \ldots, q_{L_J} f(\ddot{\beta}_{jL_J}) \right), \tag{33}$$

where the likelihood is defined as

$$f(\beta) = \exp\left(\sum_{i=1}^{N} \sum_{j=1}^{J} I[y_i = j] \eta_{ij} - \log\left(\sum_{j=1}^{J} \exp(\eta_{ij}) \right) \right), \tag{34}$$

with $\eta_{ij} = \alpha_j + x_i' \beta_j$. If outcome category j is assigned to outcome cluster l, the parameter matrix β equals

$$\ddot{\beta}_{jl} = (\beta_1, \ldots, \beta_{j-1}, \tilde{\beta}_l, \beta_{j+1}, \ldots, \beta_J)'. \tag{35}$$

A.3.5 Sample cluster probabilities q and concentration parameter λ_J

Sample the unconditional cluster probabilities for the outcome categories from $q|C, \lambda_J$ according to

$$q_1 = V_1^*, \quad q_l = (1 - V_1^*)(1 - V_2^*) \dots (1 - V_{l-1}^*)V_l^*, \text{ for } l = 2, \dots, L_J - 1,$$

where

$$V_l^* \sim \text{Beta}\left(1 + r_l, \lambda_J + \sum_{k=l+1}^{L_J} r_k\right), \quad l = 1, \dots, L_J - 1, \tag{36}$$

with r_l the number of values in C which equal l.

Sample the concentration parameter λ_J according to

$$\lambda_J | q, \eta_{J1}, \eta_{J2} \sim \text{Gamma}\left(L_J + \eta_{J1} - 1, \eta_{J2} - \sum_{l=1}^{L_J - 1} \log(1 - V_l^*)\right). \tag{37}$$

A.3.6 Sample the classification variables D

Sample the classification variables of the explanatory categories as

$$D_k | p, \alpha, \tilde{\beta}, C, y, X \sim \sum_{l=1}^{L_K} \psi_{lk} \delta_l, \tag{38}$$

for $k = 1, \dots, K_d$. The conditional cluster probability ψ_{lk} is a function of the unconditional cluster probability p_l and the data likelihood:

$$(\psi_{1k}, \dots, \psi_{L_K,k}) \propto \left(p_1 f(\ddot{\beta}_{k1}), \dots, p_{L_K} f(\ddot{\beta}_{k,L_K})\right), \tag{39}$$

where the likelihood is defined in (34), and $\ddot{\beta}_{kl}$ is defined as the parameter matrix in case explanatory category k is assigned to explanatory cluster l:

$$\ddot{\beta}_{kl} = (\gamma, \kappa_{.1}, \dots, \kappa_{.k-1}, \tilde{\kappa}_{.l}, \kappa_{.k+1}, \dots, \kappa_{.K_d}). \tag{40}$$

A.3.7 Sample cluster probabilities p and concentration parameter λ_K

Sample the unconditional cluster probabilities for the explanatory categories p in the same way as for q. Similarly, the concentration parameter for the explanatory categories λ_K is sampled in the same way as for λ_J.

A.3.8 Posterior simulation one-way clustering

For one-way clustering 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 3, and ignore Step 6 and 7 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 Steps 4 and 5.

A.4 Predictive distribution

We simulate from the predictive distribution of y_i in iteration s of the sampler as

$$y_i^{(s)} \sim \text{Multinomial}(1, \phi_i^{(s)}), \tag{41}$$

where the probability vector $\phi_i^{(s)}$ has elements

$$\phi_{ij}^{(s)} = P(y_i^{(s)} = j|x_i) = \frac{\exp(\eta_{ij}^{(s)})}{\sum_{j=1}^{J} \exp(\eta_{ij}^{(s)})}, \quad \eta_{ij}^{(s)} = \alpha_j^{(s)} + \tilde{\gamma}_{C_j^{(s)}}^{(s)\prime} w_i + \sum_{k=1}^{K_d} \tilde{\kappa}_{C_j^{(s)}, D_k^{(s)}}^{(s)} d_{ik},$$

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

B Numerical experiments

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 one-way mixture models that cluster over outcome categories or explanatory

categories, and a standard multinomial logit model. We consider a data generating process along the dimensions of the empirical application. Next, we study the sensitivity of the results against an increase in the prior belief about the number of unique parameter values, increasing model parameter prior variance, and the setting in which the number of parameters is larger than the number of observations.

B.1 Set-up

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

$$P(y_i = j|x_i) = \frac{\exp(\eta_{ij})}{\sum_{j=1}^{J} \exp(\eta_{ij})}, \text{ with } \eta_{ij} = \alpha_j + \gamma_j' w_i + \kappa_j' d_i, \tag{42}$$

with $j = 1, \ldots, J$, and $i = 1, \ldots, N + 10,000$ where the final $10,000$ observations are used for out-of-sample analysis. The vector w_i includes four standard normally distributed variables. The categorical dummies are drawn from a multinomial distribution

$$(d_{i1}, \ldots, d_{i,K_d}, d_{i,K_d+1}) \sim \text{Multinomial}\left(\frac{p_{d_i}}{K_d}, \ldots, \frac{p_{d_i}}{K_d}, 1 - p_{d_i}\right), \tag{43}$$

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

We follow the dimensions of the empirical application and set the number of outcome categories to $J = 50$ and the number of explanatory categories to $K_d = 10$. The intercepts have the values $\alpha_1 = 0$, and $\alpha_j \sim U[-1, 1]$ sampled from a uniform distribution for $j = 2, \ldots, 50$. The outcome and explanatory categories are both clustered into five

groups, with model parameter values equal to

$$\tilde{\beta}_l = \begin{pmatrix} \tilde{\gamma}_l \\ \tilde{\kappa}_{l,1} \\ \tilde{\kappa}_{l,2} \\ \tilde{\kappa}_{l,3} \\ \tilde{\kappa}_{l,4} \\ \tilde{\kappa}_{l,5} \end{pmatrix} = \begin{cases} (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)' & \text{if } l = 1, \\ (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad -2 \quad -2 \quad 2 \quad 2)' & \text{if } l = 2, \\ (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad -1 \quad 1 \quad -1 \quad 1)' & \text{if } l = 3, \\ (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad -1 \quad 1 \quad -1)' & \text{if } l = 4, \\ (1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 2 \quad 2 \quad -2 \quad -2)' & \text{if } l = 5, \end{cases} \tag{44}$$

where $\beta_j = (\tilde{\gamma}'_{C_j}, \tilde{\kappa}_{C_j,D_1}, \ldots, \tilde{\kappa}_{C_j,D_{10}})'$ with

$$C_j = \begin{cases} 1 & \text{if } 1 \leq j \leq 10, \\ 2 & \text{if } 11 \leq j \leq 20, \\ 3 & \text{if } 21 \leq j \leq 30, \\ 4 & \text{if } 31 \leq j \leq 40, \\ 5 & \text{if } 41 \leq j \leq 50, \end{cases} \quad \text{and} \quad D_k = \begin{cases} 1 & \text{if } k = 1, 2, \\ 2 & \text{if } k = 3, 4, \\ 3 & \text{if } k = 5, 6, \\ 4 & \text{if } k = 7, 8, \\ 5 & \text{if } k = 9, 10. \end{cases} \tag{45}$$

Table 1 specifies the dimensions of the simulated data and the prior distributions of the model parameters for four different experiments. Experiment 1 estimates the models on $N = 4000$ observations with the settings as discussed in Section 4. Experiments 2-4 are designed to examine the sensitivity against the settings in experiment 1. Experiment 2 sets the prior distributions of the concentration parameters according to the prior belief that the mode of unique parameter values across outcome categories equals 20 and across explanatory categories equals 8, instead of respectively 15 and 5 in Experiment 1. Experiment 3 increases the model parameter prior variance from $\sigma_\beta^2 = 1$ to $\sigma_\beta^2 = 2$. Experiment 4 lets the number of parameters (735) exceed the number of observations $N = 400$ instead of $N = 4000$.

Posterior results are based on 1,000,000 iterations of the Gibbs sampler, from which the first 500,000 are discarded and we use a thinning value of 50.

Table 1: Settings numerical experiments

Experiment	Prior distribution λ_J	Prior distribution λ_K	σ^2_β	N
1	Gamma$(7.15 \times 20, 20)$	Gamma$(3.47 \times 1, 1)$	1	4000
2	Gamma$(12.24 \times 20, 20)$	Gamma$(15.10 \times 1, 1)$	1	4000
3	Gamma$(7.15 \times 20, 20)$	Gamma$(3.47 \times 1, 1)$	2	4000
4	Gamma$(7.15 \times 20, 20)$	Gamma$(3.47 \times 1, 1)$	1	400

This table shows the differences between the numerical experiments in Appendix B. Experiment 1 in the first row is the standard setup. The remaining rows show the settings in the other experiments, with the differences between the experiments and experiment 1 indicated by the gray cells.

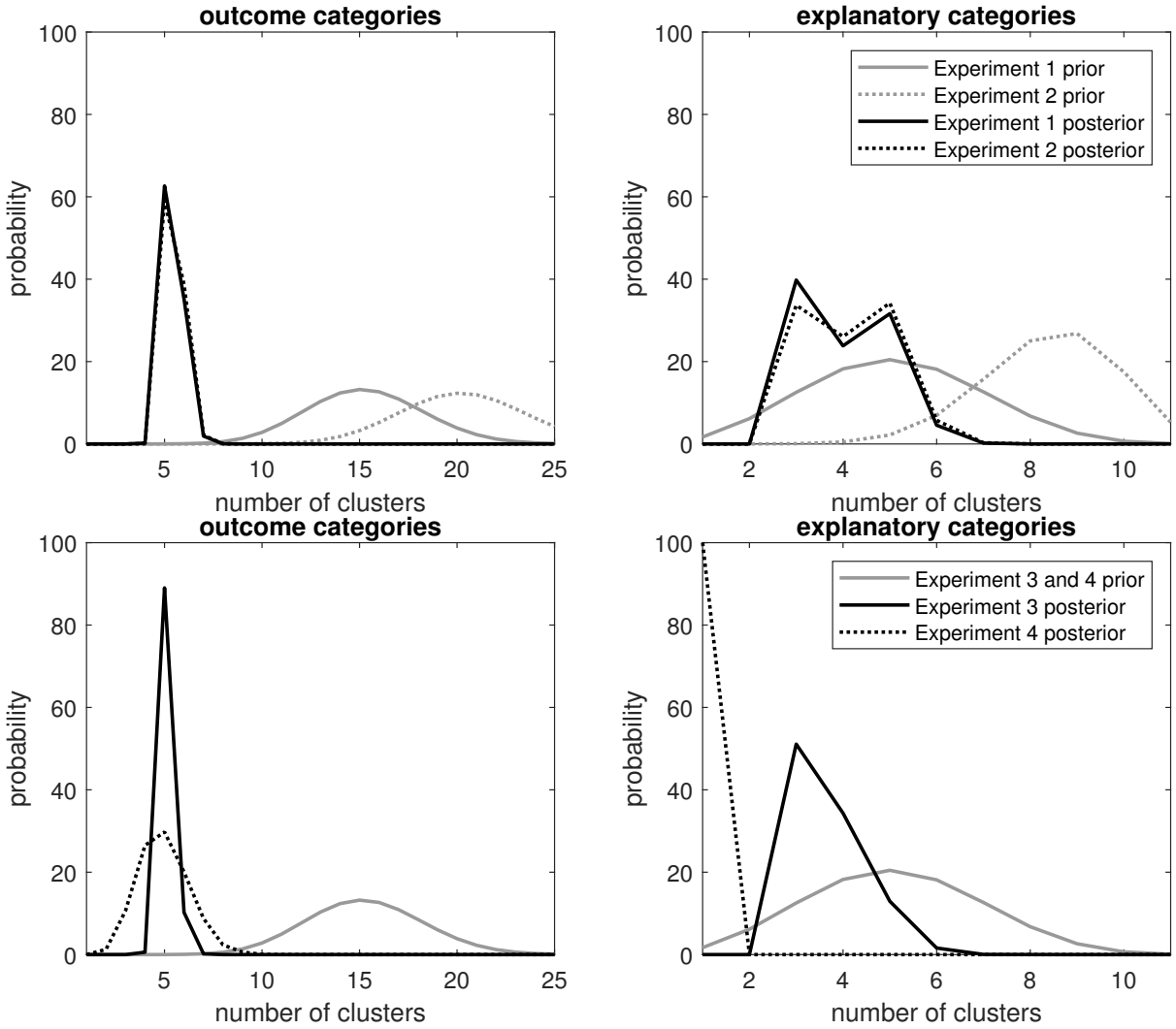
B.2 Results

Figure 1 shows the posterior distributions of the number of unique parameter values across outcome categories and explanatory categories in the two-way mixture model, for experiment 1-4. The model substantially reduces the number of model parameters in experiment 1: The posterior number of clusters for the fifty outcome categories is tightly concentrated around five. The posterior distribution for the explanatory categories is more diffuse, but still puts a substantial probability mass on the correct number of clusters.

The posterior distributions in Figure 1 are not very sensitive to an increase in the prior belief on the number of clusters, or an increase in the prior model variance. The differences between the prior distributions on the number of clusters in experiment 1 and 2 is substantial, with way more probability mass on a large number of clusters in experiment 2. However, the posterior distribution in experiment 2 only slightly moves to the right compared to the posterior distribution in experiment 1. The posterior distributions of the number of clusters in experiment 3 are also similar to experiment 1. Both for the outcome and explanatory categories, the distributions are less diffuse in experiment 3, and the posterior for the explanatory categories puts more probability mass on small numbers of clusters. Experiment 4 decreases the number of observations from 4000 in experiment 1-3 to 400. This results in a more diffuse posterior across outcome categories, and a posterior that does not find any variation across explanatory categories.

Table 2 shows the in- and out-of-sample log-score and hit-rate for experiments 1-4.

Figure 1: Distribution of the number of unique parameter values



This figure shows the prior distributions (gray lines) and posterior distribution in the two-way mixture model (black lines) over the number of unique parameter values over outcome categories (left panel) and explanatory categories (right panel). Section A.2 discusses these distributions.

The mixture models are compared to a standard multinomial logit 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. Two-way clustering improves the out-of-sample log-score and hit-rate relative to the standard multinomial logit model in each experiment. Moreover, two-way clustering also improves on these metrics relative to one-way clustering across outcomes or dummies in each experiment. Two-way clustering also performs well in-sample, outperforming standard MNL in experiments 1-3, but is not improving in-sample upon the standard MNL in experiment 4.

Table 2: Log-score and hit-rate for numerical experiments

		clustering			standard	
sample	metric	two-way	outcomes	dummies	MNL	naive
		Experiment 1: settings empirical application				
in	log-score	-3.390	-3.386	-3.405	-3.419	-3.758
in	hit-rate	0.085	0.086	0.085	0.084	0.054
out	log-score	-3.408	-3.422	-3.475	-3.512	-3.763
out	hit-rate	0.090	0.090	0.085	0.081	0.051
		Experiment 2: concentration parameters				
in	log-score	-3.384	-3.391	-3.406	-3.419	-3.758
in	hit-rate	0.087	0.087	0.085	0.084	0.054
out	log-score	-3.403	-3.422	-3.478	-3.512	-3.763
out	hit-rate	0.091	0.088	0.086	0.081	0.051
		Experiment 3: model parameters				
in	log-score	-3.394	-3.421	-3.410	-3.413	-3.758
in	hit-rate	0.090	0.088	0.084	0.081	0.054
out	log-score	-3.413	-3.448	-3.491	-3.525	-3.763
out	hit-rate	0.090	0.088	0.084	0.079	0.051
		Experiment 4: number of observations				
in	log-score	-3.529	-3.579	-3.538	-3.446	-3.726
in	hit-rate	0.078	0.065	0.068	0.080	0.058
out	log-score	-3.653	-3.730	-3.793	-3.760	-3.807
out	hit-rate	0.065	0.044	0.053	0.059	0.051

This table shows in-sample and out-of-sample log-score and hit-rate for different experiments, as defined in (16) and (17), respectively.

This may be explained by Figure 1. The posterior distributions of the number of clusters are similar across experiments 1-3. In experiment 4, the posterior of the two-way mixture model does not find any variation across explanatory categories. Table 2 shows that this mainly affects the in-sample model fit.

Table 3 shows the mean squared error (MSE) of the posterior parameter draws and the interquartile range (IQR) of the posterior parameter distributions for the different models. Two-way clustering improves the MSE and IQR in experiment 1-3 compared to one-way clustering and the standard MNL model. We find that two-way clustering increases the MSE and decreases the IQR compared to standard MNL in experiment 4. This suggests that a decrease in the variance of the parameter estimates is at the expense

Table 3: Mean squared error and interquartile range for numerical experiments

		clustering			standard
experiment	metric	two-way	outcomes	dummies	MNL
1	MSE	0.264	0.768	0.475	0.672
1	IQR	0.342	0.515	0.533	0.811
2	MSE	0.258	0.807	0.476	0.672
2	IQR	0.352	0.511	0.532	0.811
3	MSE	0.325	1.360	0.444	0.661
3	IQR	0.239	0.513	0.616	0.999
4	MSE	1.185	2.886	1.161	1.067
4	IQR	0.129	0.730	0.283	1.157

This table shows mean squared error (MSE) of the posterior draws and the interquartile range (IQR) of the posterior parameter distributions, averaged over all model parameters.

of an increase in bias.

The differences in performance of the two-way mixture model across experiments, follow the posterior distributions of the number of clusters in Figure 1. These distributions are similar in experiment 1-2, as is the model fit and accuracy of the parameter estimates as evaluated in Tables 2 and 3. The fact that the distributions are less diffuse in experiment 3 is reflected in a lower IQR. The posterior distribution of the number of clusters across explanatory categories in experiment 3 puts more probability mass on a smaller number of clusters than five. This results in biased parameter estimates, which is captured by an increase in MSE. Finally, the posterior distributions in Figure 1 for experiment 4 have more uncertainty around the number of clusters across outcome categories, and do not include the correct number of clusters across explanatory categories. This results in worse predictive performance and a large MSE of the parameter estimates.

C Mixed logit model

Let y_{it} be an observable random categorical variable, such that $y_{it} \in \{1, 2, \ldots, J\}$, with J the number of choice alternatives, $i = 1, \ldots, N$, with N the number of individuals, and $t = 1, \ldots, T$, with T the number of time periods. Let x_{it} be a K_x-dimensional vector

with explanatory variables that vary across individuals, and z_{itj} an explanatory variable that varies across individuals and choice alternatives. The probability that individual i in time period t chooses alternative j is

$$P(y_{it} = j|x_{it}, z_{itj}) = \frac{\exp(\eta_{itj})}{\sum_{j=1}^{J} \exp(\eta_{itj})}, \tag{46}$$

where η_{itj} is a linear function of parameters for all $j = 1, \ldots, J$,

$$\eta_{itj} = \alpha_j + x'_{it}\beta_j + z'_{itj}\nu_{ij}, \tag{47}$$

with alternative-specific intercept α_j, K_x-dimensional coefficient vector β_j, and random coefficients ν_{ij} with $\nu_{i1} = 0$ and

$$\nu_i = (\nu_{i2}, \ldots, \nu_{iJ})' \sim N(u, Q), \tag{48}$$

where u is a $(J-1)$-dimensional mean vector and Q a $(J-1)$-dimensional covariance matrix.

The multinomial logit model in (1) and (2) sets z_{itj} equal to zero. As a result, there is no correlation in the utilities across alternatives and the IIA property holds. Nonzero z_{itj} with a diagonal covariance matrix Q allow for restrictive substitution patterns and hence do not impose IIA. If z_{itj} is nonzero and the nondiagonal elements of Q are nonzero, the utilities are allowed to be correlated across alternatives which allows for general substitution patterns. Since J is large in our case, we model Q as a factor covariance matrix: $Q = \Lambda\Lambda' + \Psi$, with a $(J-1)$-dimensional vector of factor loadings Λ and a $(J-1)$-dimensional diagonal covariance matrix Ψ. The prior distributions for the additional parameters are $u \sim N(0_{J-1}, \sigma_u^2 I_{J-1})$, $\Lambda \sim N(0_{J-1}, \sigma_\Lambda^2 I_{J-1})$, and $\Psi \sim$ Inverse-Gamma(a_Ψ, b_Ψ).

Similar as in the multinomial logit model, (47) can be rewritten to

$$\eta_{itj} = \alpha_j + w'_{it}\gamma_j + \sum_{k=1}^{K_d} \kappa_{jk} d_{itk} + z'_{itj}\nu_{ij}, \tag{49}$$

and the two-way Dirichlet process prior for γ_j and κ_{jk} in (13) can be used.

C.1 Posterior simulation

The sampling steps for the mixed logit model with a two-way Dirichlet process prior are similar as in Appendix A.3, with three main differences. First, on top of the initialization steps in Appendix A.3, set $u = 0_{J-1}$, $\Psi = I_{J-1}$, $\Lambda = 0_{J-1}$, and $\nu_{ij} = 0$. Second, the latent variables ω_{itj} are now sampled as

$$\omega_{itj}|\alpha_j, \tilde{\beta}, C, D, \nu_{ij}, x_{it}, z_{itj} \sim \text{PG}(1, \eta_{itj}), \tag{50}$$

for all $i = 1, \ldots, N$, $t = 1, \ldots, T$, and $j = 2, \ldots, J$.

Third, the coefficients are sampled as follows. Define $y = (y_1', \ldots, y_N')'$ and $X = (x_1', \ldots, x_N')'$, with $y_i = (y_{i1}, \ldots, y_{iT})'$ and $x_i = (x_{i1}', \ldots, x_{iT}')'$. Given $D = (D_1, \ldots, D_{K_d})$, define the K^*-dimensional vector $x_{it}^* = (w_{it}', d_{it}^{*\prime})'$, with $d_{it}^* = (\sum_{k=1}^{K_d} I[D_k = D_1^*]d_{itk}, \ldots, \sum_{k=1}^{K_d} I[D_k = D_{m_d}^*]d_{itk})'$ where $D^* = \{D_1^*, \ldots, D_{m_d}^*\}$ denote the current m_d unique values of D. The rows of the $NT \times K_x^*$ regressor matrix X^* equal the $x_{it}^{*\prime}$. For all nonempty outcome clusters l,

$$\tilde{\beta}_l|C, D, \alpha, \nu, \sigma_\beta^2, \omega, y, X, Z \sim N(b_l, B_l), \tag{51}$$

with $b_l = B_l X^{*\prime} \sum_{j=2}^J 1[C_j = l](\zeta_j + \omega_j \odot m_j)$ and $B_l = (\sum_{j=2}^J 1[C_j = l] X^{*\prime}\text{diag}(\omega_j)X^* + V_b)^{-1}$, where the elements of the NT-dimensional vector ζ_j equal $1[y_{it} = j] - 0.5$, the elements of the NT-dimensional vector ω_j equal ω_{itj}, and the elements of the NT-dimensional vector m_j equal $\log \sum_{k \neq j} \exp \eta_{itk} - \alpha_j - z_{itj}'\nu_{ij}$. The NT-dimensional diagonal matrix $\text{diag}(\omega_j)$ has the elements in ω_j on the diagonal, and the K_x^*-dimensional diagonal matrix has σ_β^{-2} on the diagonal. The coefficients corresponding to empty outcome clusters and empty explanatory clusters are sampled from the base distribution as in Appendix A.3.

The alternative-specific intercepts α_j are sampled as

$$\alpha_j|C, D, \tilde{\beta}, \nu, \sigma_\alpha^2, \omega, y, X, Z \sim N(a_j, A_j), \tag{52}$$

with $a_j = A_j(\sum_{it} \zeta_{itj} + \omega_{itj}(\log \sum_{k \neq j} \exp \eta_{itk} - x'_{it}\beta_j - z'_{itj}\nu_{ij}))$ and $A_j = (\sum_{it} \omega_{itj} + \sigma_\alpha^{-2})^{-1}$. The random coefficients ν_{ij} are sampled as

$$\nu_i|C, D, \tilde{\beta}, \alpha, u, Q, \omega, y, X, z \sim N(v_i, V_i), \tag{53}$$

where $v_i = V_i(\text{diag}(\{\sum_t z_{itj}(\zeta_{itj} + \omega_{itj}(\log \sum_{k \neq j} \exp \eta_{itk} - x'_{it}\beta_j - \alpha_j))\}_{j=2}^J) + Q^{-1}u)$ and $V_i = (\text{diag}(\{\sum_t \omega_{itj} z_{itj}^2\}_{j=2}^J) + Q^{-1})^{-1}$. The mean vector of the random coefficient is samples as

$$u|\nu, Q, \sigma_u^2 \sim N\left(\frac{m}{M}, \frac{1}{M}Q\right), \tag{54}$$

where the elements of the $(J-1)$-dimensional vector m equal $\sum_i \nu_{ij}$ and $M = N + \sigma_u^{-2}$. Finally, Λ and Ψ in $Q = \Lambda\Lambda' + \Psi$ are sampled in a factor analysis model as discussed in, for instance, Section 8.3.2 of Greenberg (2012).

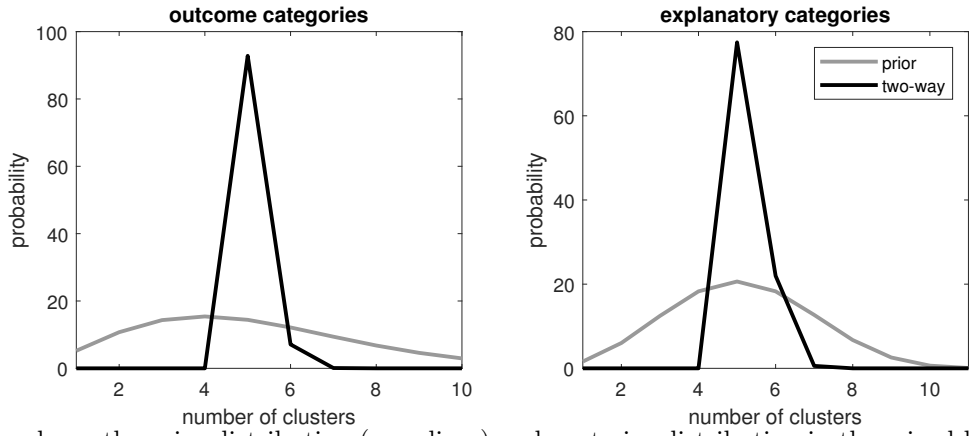
The classification variables, cluster probabilities, and concentration parameters are sampled along the lines of the steps in Appendix A.3.

C.2 Numerical experiment

The choice data are generated in the same way as in Appendix B, with two exceptions. First, $N = 1,000$ and $T = 5$. Second, (47) also includes z_{itj} and ν_{ij}, where z_{itj} is a scalar generated from a standard normal distribution. The random coefficients are generated from (48) with u generated from a standard normal distribution, the diagonal elements of Q equal 1, and the off-diagonal elements $Q_{jk} = (j-1)(k-1)/(J-1)^2$.

The prior distributions of the concentration parameters are set according to the prior belief that the mode of unique parameter values across outcome and explanatory categories equals 5, with the truncation levels equal to 10 and 11, respectively:

Figure 2: Distribution of the number of unique parameter values in the mixed logit



This figure shows the prior distribution (gray lines) and posterior distribution in the mixed logit model (black lines) over the number of unique parameter values over outcome categories (left panel) and explanatory categories (right panel).

$\lambda_J \sim \text{Gamma}(1.31 \times 2, 2)$ and $\lambda_K \sim \text{Gamma}(3.47 \times 1, 1)$. The model parameter prior variance equals $\sigma_\alpha^2 = \sigma_\beta^2 = 1$. The prior distributions of the additional parameters equal $u \sim N(0, I)$, $\Lambda \sim N(0, I)$, and $\Psi \sim \text{Inverse-Gamma}(5, 1)$.

Posterior results are based on 1,000,000 iterations of the Gibbs sampler, from which the first 500,000 are discarded and we use a thinning value of 50.

C.3 Results

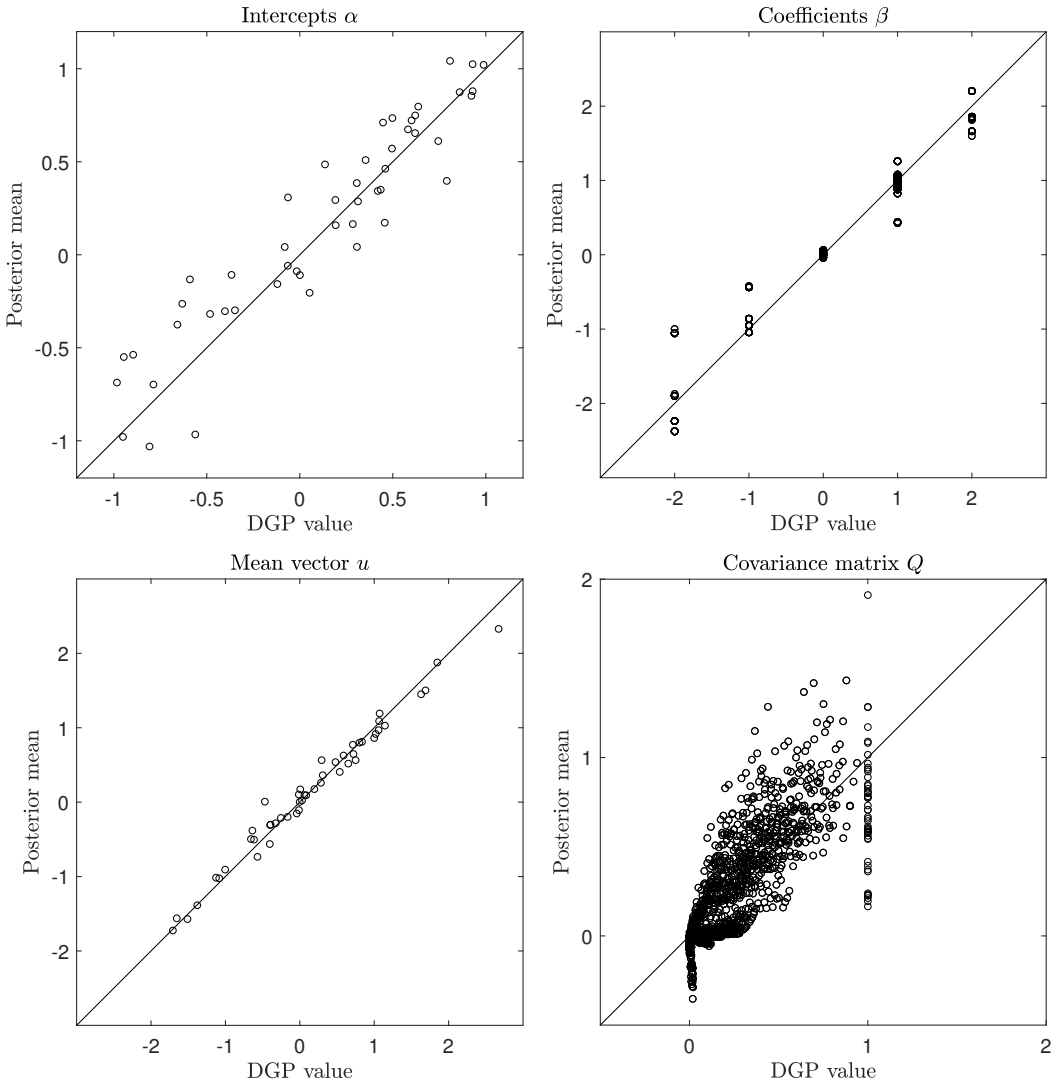
We assess the convergence of the MCMC sampler using two different diagnostics. First, we test for convergence of the sampler by the Geweke (1992) t-test for the null hypothesis of equality of the means computed from the first 20 percent and the last 40 percent of the sample draws. We compute the variances 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. We reject for 12.4%, 5.3%, and 1.3% of all estimated parameters in α, β, u, and Q the null-hypothesis, on a significance level of 10%, 5%, and 1% respectively.

Second, we analyze the 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 effective sample size for a parameter equals the number of samples $S = 10,000$ divided by

the corresponding inefficiency factor. The median, mean, and minimum effective sample size equal 2,535, 3,389, and 353, respectively.

Next we analyse the results. Figure 2 shows the posterior distributions of the number of unique parameter values across outcome categories and explanatory categories in the mixed logit model. The model substantially reduces the number of model parameters, with most probability mass on the number of clusters in the data generating process for both the outcome and explanatory categories.

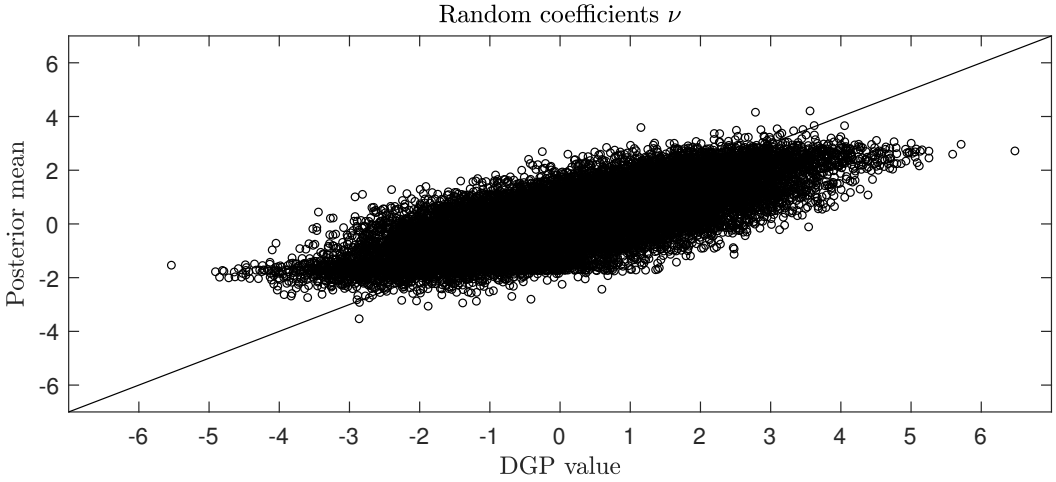
Figure 3: Posterior means of the parameters in the mixed logit



The panels in this figure correspond to the posterior means of the alternative-specific intercepts α_j, the coefficients β_{jk}, the means of the random coefficients u_j, and the elements of the covariance matrix Q of the random coefficients. The panels show the posterior means on the y-axis and the values in the data generating process on the x-axis.

Figure 3 compares the parameter values in the data generating process against the cor-

Figure 4: Posterior means of the random coefficients in the mixed logit



This figure shows the random coefficients ν_{ij}. The panels show the posterior means on the y-axis and the values in the data generating process on the x-axis.

responding posterior mean estimates. The closer the circles lie to the 45 degree diagonal line, the closer the posterior means are to the true parameter values. The alternative-specific intercepts, the coefficients, and the means of the random coefficients are scattered around the 45 degree line. The zero covariances and the variances in the covariance matrix of the random coefficients seems to be slightly downward biased, which may be explained by the fact that accurately estimating such a high-dimensional covariance matrix is challenging (Geweke et al., 1994). The random coefficients in Figure 4 are close to the 45 degree line, but especially larger values show some shrinkage towards zero.

D Empirical application

D.1 Overview categorical dependent variable

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

1	France	18	Eastern Europe	35	Western Asia
2	Iceland	19	Portugal	36	Southern Asia
3	Norway	20	Spain	37	China
4	Sweden /Finland	21	Italy	38	Eastern Asia
5	Denmark	22	Malta	39	Indonesia
6	Ireland	23	Croatia	40	Thailand
7	United Kingdom	24	Greece	41	Southeastern Asia
8	Belgium	25	Southern Europe	42	Australia/ New Zealand
9	Luxembourg	26	Morocco	43	Canada
10	Germany	27	Tunisia	44	United States
11	Switzerland	28	Egypt	45	Netherlands Antilles
12	Austria	29	Eastern Africa	46	Caribbean
13	Poland	30	West Africa	47	Mexico
14	Czech Republic	31	Southern Africa	48	Central America
15	Hungary	32	Cyprus	49	Southern America
16	Romania	33	Israel		
17	Bulgaria	34	Turkey		

Table 4: Descriptive statistics holiday spells per holiday destination

		total				estimation			
j	destinations	min	median	mean	max	min	median	mean	max
1	France	2	13	14.213	85	8	15	17.554	85
2	Iceland	4	12	12.217	27	9	14.5	15.188	27
3	Norway	3	14	14.134	37	8	15	15.789	37
4	Sweden/Finland	3	11	15.507	56	8	17.5	20.460	56
5	Denmark	2	8	9.638	29	8	11	12.943	29
6	Ireland	2	10	9.596	19	8	11	11.889	19
7	United Kingdom	2	5	7.649	45	8	11	13.191	45
8	Belgium	2	4	5.452	44	8	9	11.348	44
9	Luxembourg	2	7	8.744	23	8	13	13.514	23
10	Germany	2	5	7.046	78	8	10	12.869	78
11	Switzerland	3	9	11.824	51	8	10	13.805	51
12	Austria	3	9	10.836	37	8	9	11.851	37
13	Poland	3	8	9.083	27	8	10	13.038	27
14	Czech Republic	3	9	10.194	33	8	13.5	13.775	33
15	Hungary	3	11	14.347	66	8	18	19.806	66
16	Romania	3	11.5	13.143	48	8	12.5	16.600	48
17	Bulgaria	4	10	9.769	15	8	11	10.636	15
18	Eastern Europe	3	10	11.464	42	8	12	14.150	42
19	Portugal	3	9	12.250	64	8	12	14.625	64
20	Spain	2	10	13.202	89	8	12	14.660	89
21	Italy	2	11	12.623	67	8	14	14.958	67
22	Malta	4	8	8.476	15	8	8	9.438	15
23	Croatia	7	16	18.041	71	8	16	18.194	71
24	Greece	4	11	12.061	55	8	11	12.284	55
25	Southern Europe	8	12	17.258	66	8	12	17.258	66
26	Morocco	4	9	11.000	26	8	9.5	12.773	26
27	Tunisia	8	8.5	12.688	41	8	8.5	12.688	41
28	Egypt	7	9	10.729	17	8	9	10.793	17
29	Eastern Africa	5	17.5	18.182	47	9	19	18.810	47
30	West Africa	7	14	12.621	22	8	14.5	12.821	22
31	Southern Africa	5	21	19.375	38	9	21	20.000	38
32	Cyprus	8	10	11.412	20	8	10	11.412	20
33	Israel	8	13	15.818	29	8	13	15.818	29
34	Turkey	4	9.5	10.808	72	8	10	11.192	72
35	Western Asia	2	8.5	8.786	20	8	9	10.632	20
36	Southern Asia	9	16.5	17.556	26	9	16.5	17.556	26
37	China	6	22	22.214	51	10	22	23.462	51
38	Eastern Asia	9	19	19.500	31	9	19	19.500	31
39	Indonesia	10	23	24.128	84	10	23	24.128	84
40	Thailand	3	18	18.870	44	10	19	19.591	44
41	Southeastern Asia	3	21	22.258	69	11	22	24.107	69
42	Australia	15	36	21.737	67	15	36	37.737	67
43	Canada	8	22	21.744	52	8	22	21.744	52
44	United States	4	16	17.260	62	8	17	18.992	62

The following table shows which countries belong to which holiday region.

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

Table 5: Frequency counts for the number of observed holidays per household

observations	1	2	3	4	5	6	7	8
frequency	2159	875	224	63	8	3	0	2

This table shows the frequency counts for the number of observed holidays per household in the 4907 observations used in the data in the empirical application.

D.2 Overview control variables

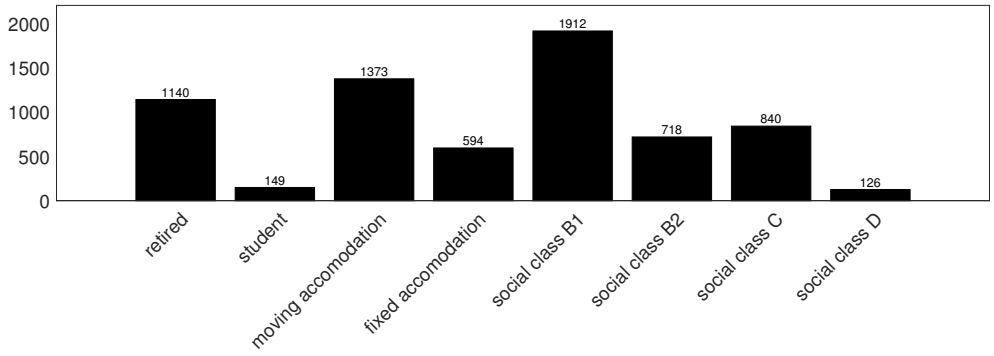
The survey contained a multiple choice question with 28 income categories. Table 6 shows the income categories, which are transformed to a continuous variable by taking the logarithm of the upper limit of each income category. Moving holiday accommodations include tents, caravans, campers, and cabin 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. Figure 5 shows the frequency counts for the binary dummies.

Table 6: 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.

Figure 5: Frequency counts dummy control variables



This figure shows the frequency counts for the explanatory control variables. The frequencies represent the number of observations that are coded as 1 in the binary dummies.

D.3 Convergence diagnostics

We assess the convergence of the MCMC sampler in the two-way mixture model in the empirical application using three different diagnostics. First, we use the Gelman–Rubin diagnostic to analyse the difference between three Markov chains with a different random initialization (Gelman and Rubin, 1992). This diagnostic compares the estimated between-chains and within-chain variances for each model parameter. Brooks and Gelman (1998) suggest that all chains have converged if the Gelman–Rubin diagnostic is smaller than 1.2 for all model parameters. The largest value we find is 1.018.

Second, we 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. We reject for 10.3%, 5.6%, and 1.2% of the 960 estimated parameters the null-hypothesis, on a significance level of 10%, 5%, and 1% respectively.

Third, we analyze the autorrelation functions of the model parameters displayed in Figure 6. We summarize the autocorrelations per model parameter with the inefficiency factors $1 + 2\sum_{f=1}^{\infty} \rho_f$, where ρ_f is the fth order autocorrelation of the chain of draws for

Figure 6: Autocorrelation functions of all model parameters

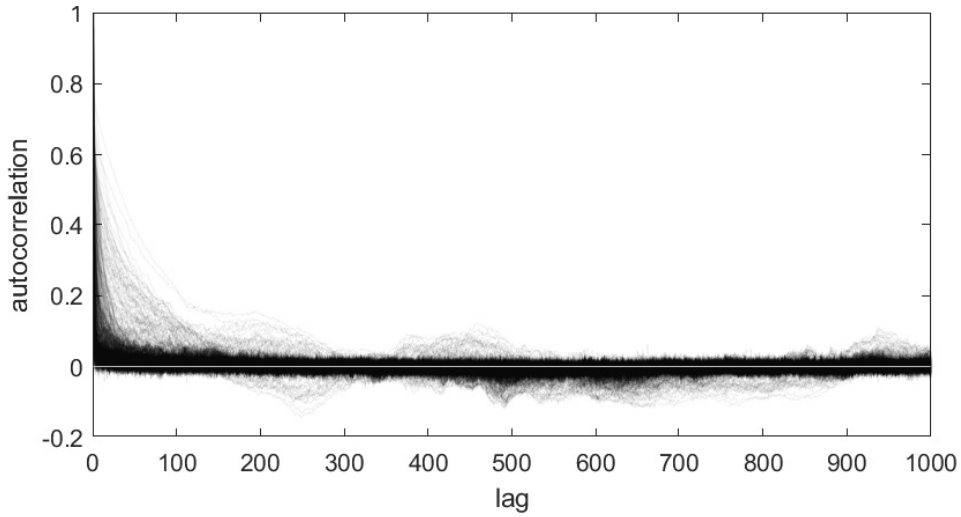
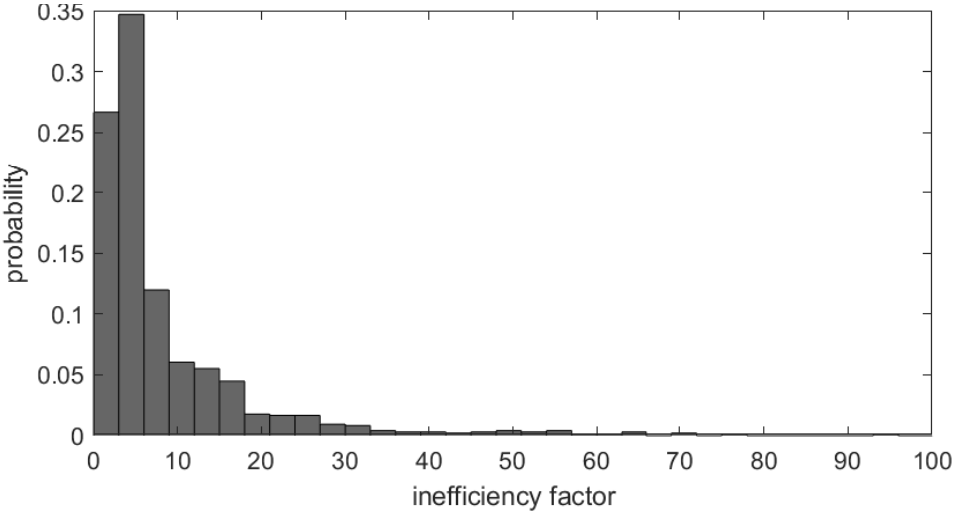


Figure 7: Inefficiency factors of all model parameters



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 effective sample size for a parameter equals the number of samples $S = 10,000$ divided by the corresponding inefficiency factor. Figure 7 shows that the effective sample size is larger than 1,000 for more than 75.9% of the model parameters.

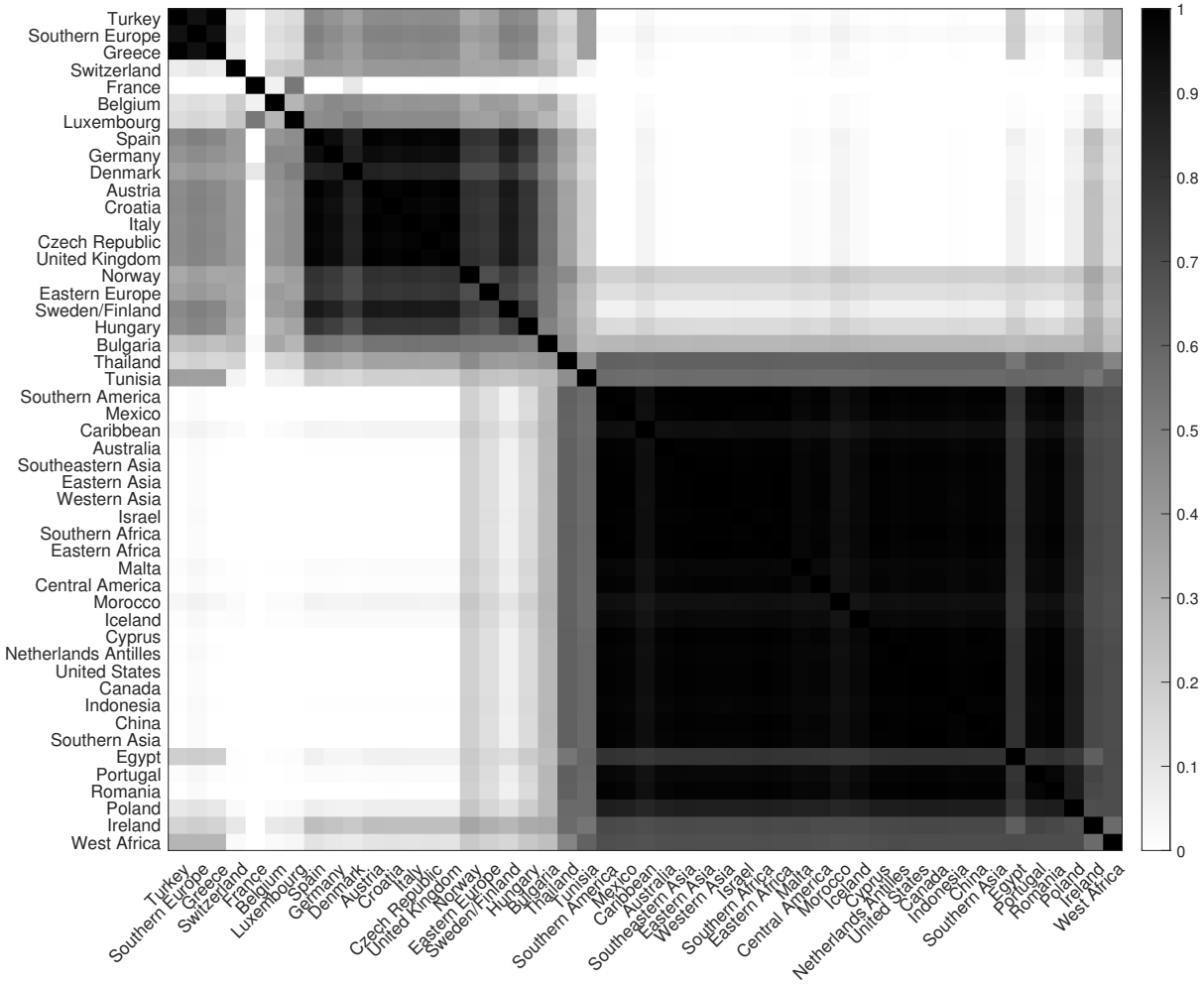
D.4 Additional empirical results

Table 7: Model evaluation p-values

		clustering		
		two-way	holiday	household
hit-rate	in	0.797	1.000	0.932
	out	0.780	0.867	0.867
log-score	in	0.000	0.000	0.000
	out	0.022	0.183	0.174

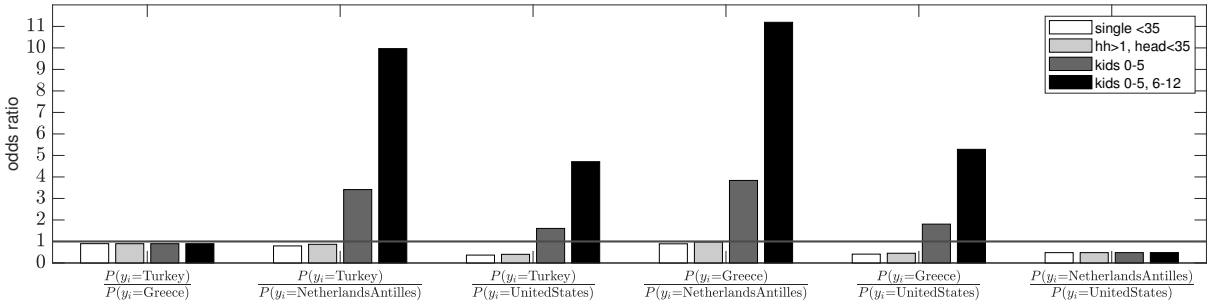
This table shows the p-values for the tests on the difference of the hit-rates and the difference of the log-scores between the method indicated by the column label and a standard MNL. See Table 2 for details.

Figure 8: One-way pairwise cluster probabilities for the holiday destinations



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 in the one-way mixture model across holiday destinations. The posterior probabilities range from zero (white) to one (black).

Figure 9: Posterior odds ratios in the one-way mixture model



This figure shows the posterior odds ratios for all combinations of the holiday destinations Turkey, Greece, Netherlands Antilles and United States, for four different household categories. The control variables are set to mean log income, not retired or student, no fixed or moving holiday accommodation, and social class A. The posterior odds are from the one-way holiday destination mixture model.