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.

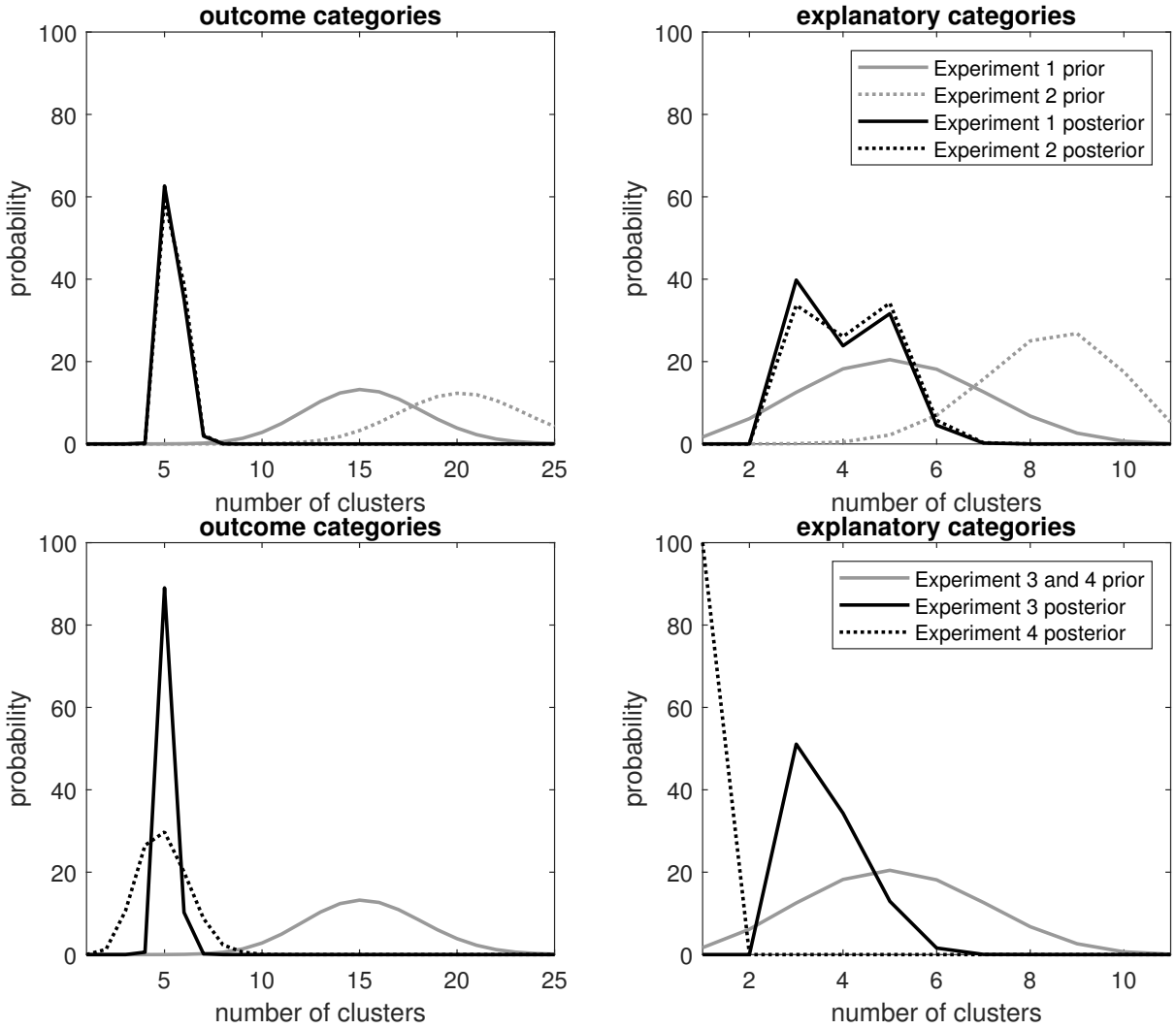
B.2 Results

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).

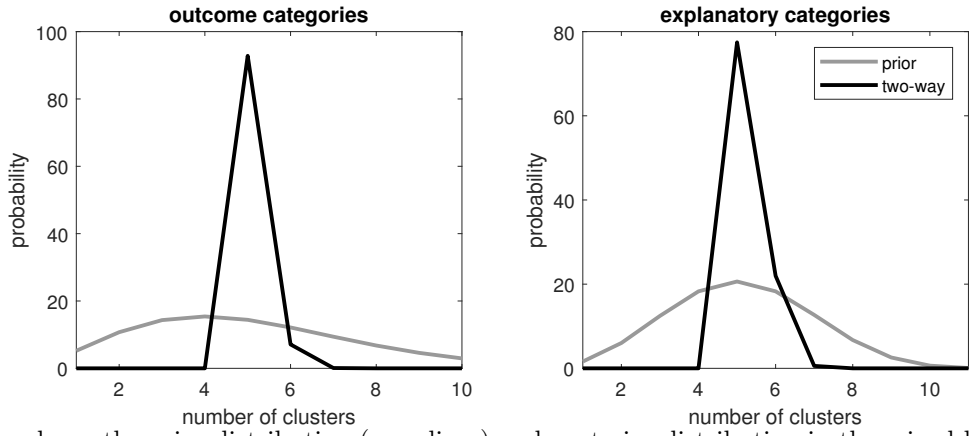
The classification variables, cluster probabilities, and concentration parameters are sampled along the lines of the steps in Appendix A.3.

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

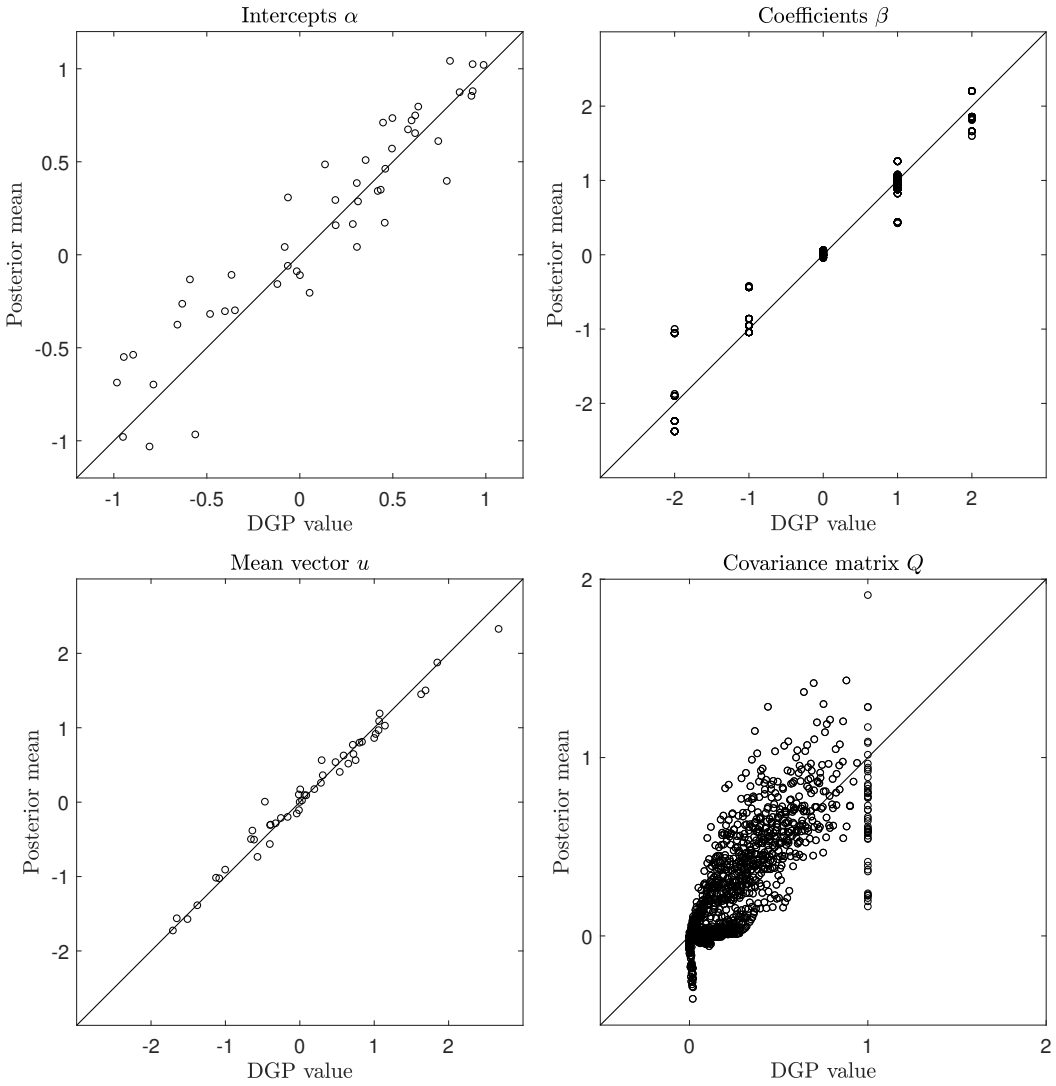
We assess the convergence of the MCMC sampler using two different diagnostics. First, we test for convergence of the sampler by the Geweke (1992) t-test for the null hypothesis of equality of the means computed from the first 20 percent and the last 40 percent of the sample draws. We compute the variances 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.

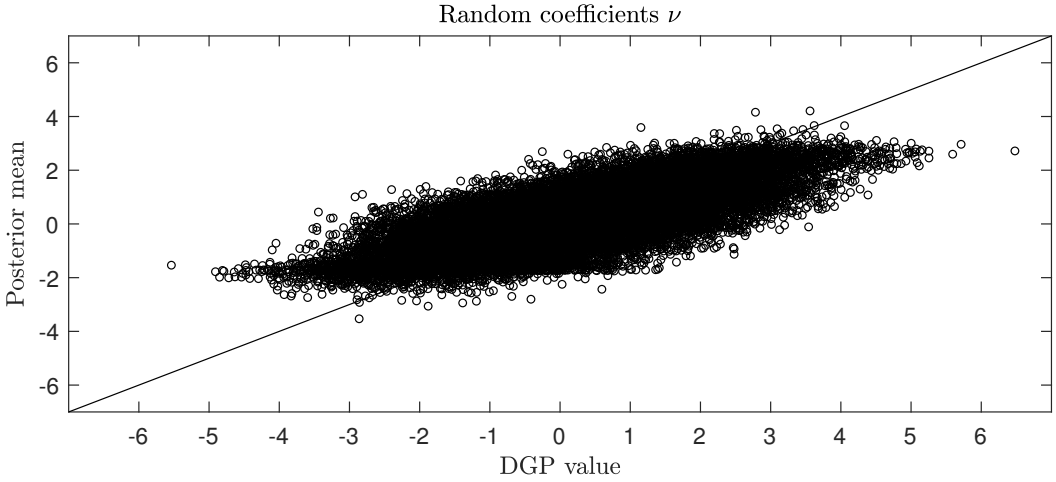
Figure 3: Posterior means of the parameters in the mixed logit



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	37.737	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
45	Netherlands Antilles	5	14.5	15.106	56	8	15	15.556	56
46	Caribbean	2	16	15.667	30	8	16	16.350	30
47	Mexico	4	14.5	15.357	24	11	15	16.231	24
48	Central America	14	21	20.615	32	14	21	20.615	32
49	Southern America	6	21	20.310	31	9	22	21.370	31
	all	2	6	8.818	89	8	13	14.835	89

This table shows the minimum, median, mean, and maximum holiday spell for each holiday destination, in the total data sample and the sample used for estimation.

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

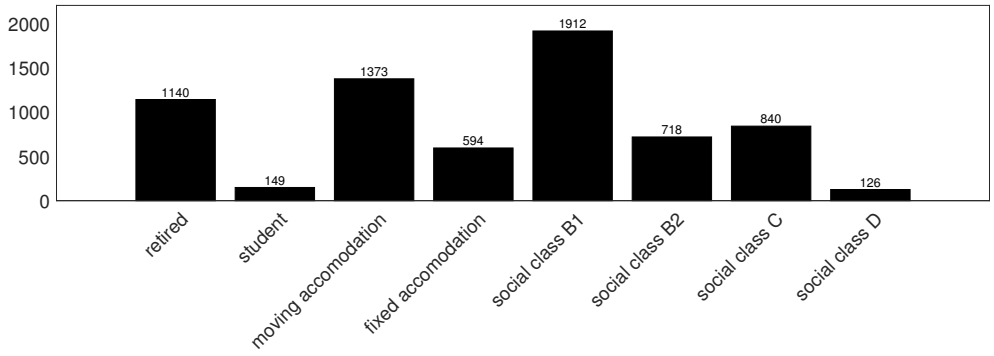
The survey contained a multiple choice question with 28 income categories. Table 6 shows the income categories, which are transformed to a continuous variable by taking the logarithm of the upper limit of each income category. Moving holiday accommodations include tents, caravans, campers, and cabin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

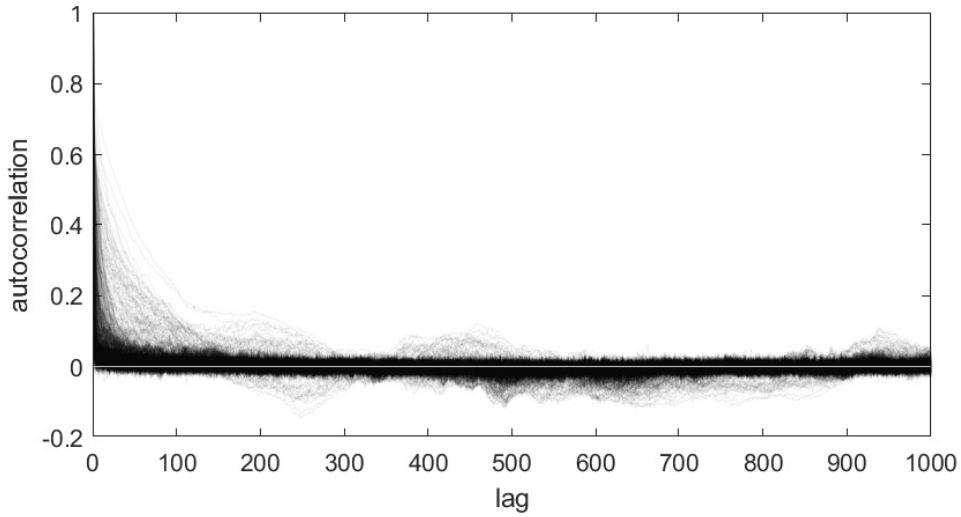
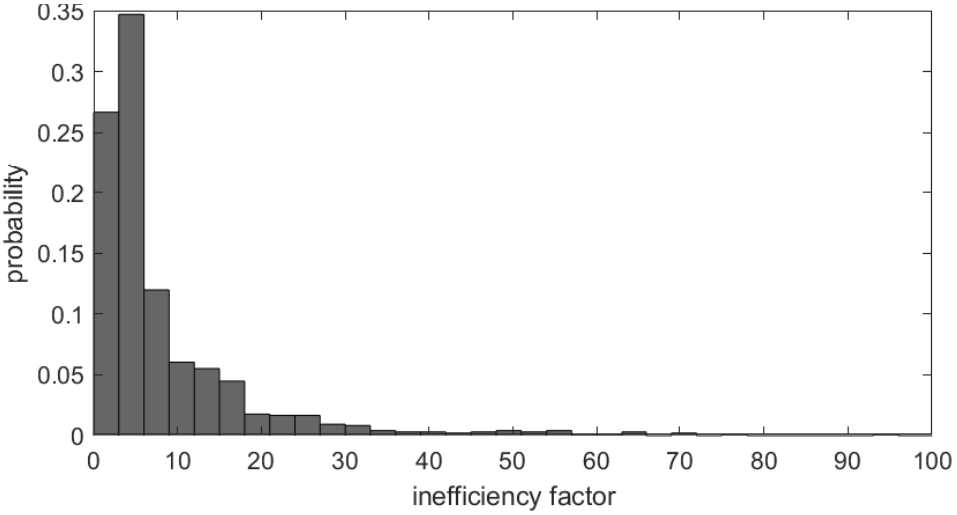


Figure 7: Inefficiency factors of all model parameters



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.

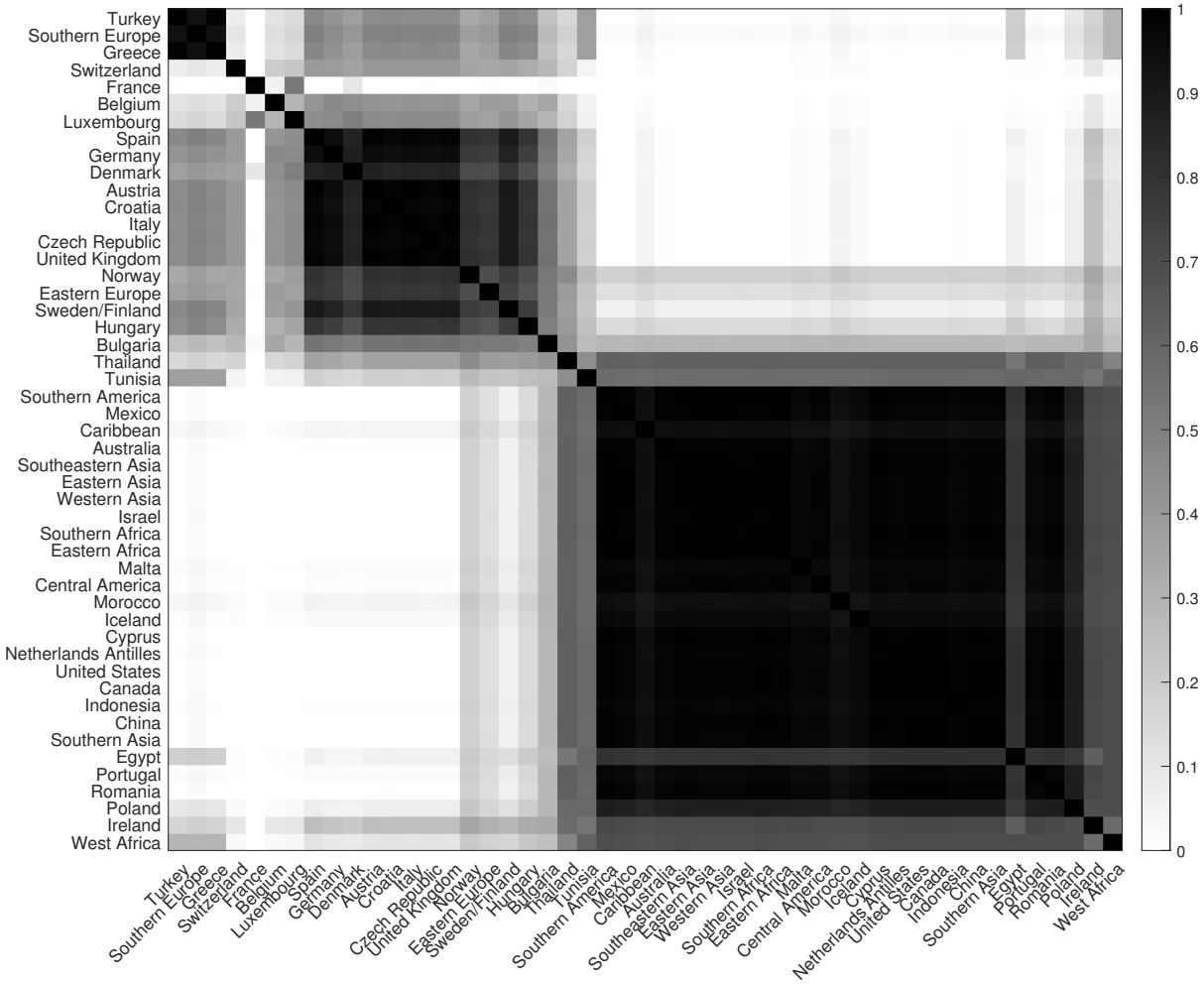
D.4 Additional empirical results

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

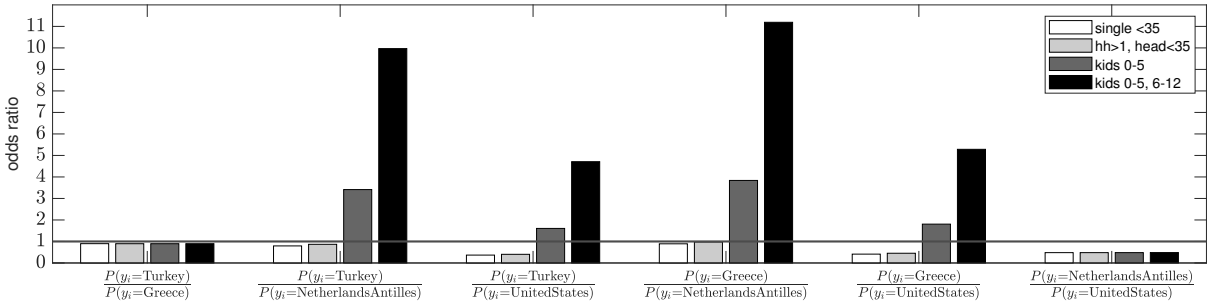
This table shows the p-values for the tests on the difference of the hit-rates and the difference of the log-scores between the method indicated by the column label and a standard MNL. See Table 2 for details.

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.