Castelein, Anoek; Fok, Dennis; Paap, Richard

Working Paper

Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach

Tinbergen Institute Discussion Paper, No. TI 2020-061/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Castelein, Anoek; Fok, Dennis; Paap, Richard (2020) : Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach, Tinbergen Institute Discussion Paper, No. TI 2020-061/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/229681

TI 2020-061/III
Tinbergen Institute Discussion Paper

Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach

Anoek Castelein[1]
Dennis Fok[1]
Richard Paap[1]

[1] Econometric Institute, Erasmus University Rotterdam

Heterogeneous variable selection in nonlinear panel data models: A semiparametric Bayesian approach*

Anoek Castelein[a] Dennis Fok[a] Richard Paap[a]

[a] Econometric Institute, Erasmus School of Economics, Erasmus University Rotterdam

September, 2020

Abstract

In this paper, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. Heterogeneous variable selection refers to the possibility that subsets of units are unaffected by certain variables. It may be present in applications as diverse as health treatments, consumer choice-making, macroeconomics, and operations research. Our method additionally allows for other forms of cross-sectional heterogeneity. We consider a two-group approach for the model's unit-specific parameters: each unit-specific parameter is either equal to zero (heterogeneous variable selection) or comes from a Dirichlet process (DP) mixture of multivariate normals (other cross-sectional heterogeneity). We develop our approach for general nonlinear panel data models, encompassing multinomial logit and probit models, poisson and negative binomial count models, exponential models, among many others. For inference, we develop an efficient Bayesian MCMC sampler. In a Monte Carlo study, we find that our approach is able to capture heterogeneous variable selection whereas a "standard" DP mixture is not. In an empirical application, we find that accounting for heterogeneous variable selection and non-normality of the continuous heterogeneity leads to an improved in-sample and out-of-sample performance and interesting insights. These findings illustrate the usefulness of our approach.

*In this paper we make use of data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands).

Keywords: individualized variable selection, Dirichlet process, stochastic search, heterogeneity, attribute non-attendance, feature selection

1. Introduction

Many panel datasets contain information on a large number of cross-sectional units with relatively little information per unit. Such datasets contain too little information to accurately estimate a separate model per unit: estimation inefficiency and overfitting would become problematic. Performing variable selection at the unit-level is therefore not straightforward. Instead, models are used that share information across units. To this end, unit-specific parameters in the model are often shrunk using an underlying population distribution shared across units. Many such distributions have been proposed: continuous distributions such as the multivariate normal or log-normal, finite mixtures of discrete or continuous distributions, and 'infinite' mixtures using a Dirichlet process.

In practice, these distributions cannot sufficiently accommodate heterogeneous variable selection on top of other cross-sectional heterogeneity. Heterogeneous variable selection refers to the possibility that subsets of units may be unaffected by certain variables. This is relevant for many applications. For example, in choice situations, groups of individuals may have no preference for or may ignore a certain product attribute when making their decisions. In macroeconomics, unemployment rates in different countries may be differentially affected or unaffected by certain macroeconomic variables. In operations research, the interarrival times of buses or the amount of garbage in bins could differentially depend or not depend on variables as temperature, holidays, or traffic conditions.

We use the term *variable selection* to denote that some units assign no weight to certain variables. Hence, variable selection is part of the data generating process. This is different from the context where variable selection refers to a researcher determining which variables should be selected in a model, also known as *model selection*. Instead of using variable *selection*, other appropriate terms are variable *importance* or variable *relevance* to indicate that for some units, certain variables may be unimportant or irrelevant.

While the literature on modeling heterogeneous responses is extensive, very few ap-

proaches have been proposed that accommodate heterogeneous variable selection. That is, the underlying population distribution to which the unit-specific parameters are shrunk, generally does not allow for groups of units to assign no weight to certain variables. Theoretically, heterogeneous variable selection can be captured when the underlying distribution is discrete, such as with a latent class approach. A discrete distribution allows the unit-specific parameters to be equal to one of multiple multivariate discrete outcomes, of which some outcomes may have certain parameters equal to zero. Practically, such a model is infeasible as the discrete distribution would need 2^K possible outcomes to capture all combinations of variable selection, where K is the number of explanatory variables. If, additionally, richer forms of heterogeneity should be allowed for, a multitude of these 2^K outcomes is needed.[1] In models with continuous heterogeneity it is even more problematic to accommodate heterogeneous variable selection, as the continuous heterogeneity distribution cannot have substantial mass at zero unless the variance of the distribution is very close to zero.

A number of papers have proposed approaches to accommodate heterogeneous variable selection. They have done so for multivariate linear models (Kim et al., 2009, Tang et al., 2020), multivariate binary probit models (Kim et al., 2018), and multinomial logit models (Gilbride et al., 2006, Scarpa et al., 2009, Hensher and Greene, 2010 Hole, 2011, Campbell et al., 2011, Hess et al., 2013, Hole et al., 2013, Collins et al., 2013, Hensher et al., 2013). Few of these papers use a Bayesian approach (Gilbride et al., 2006, Kim et al., 2009, Kim et al., 2018). The papers that use a frequentist approach have strong limitations: when allowing for flexible forms of cross-sectional heterogeneity next to heterogeneous variable selection, the developed models are susceptible to overfitting as the number of parameters quickly grows large relative to the number of observations. Furthermore, the computation time for estimation grows rapidly when the number of variables gets larger, due to the likelihood function containing 2^K terms and, in case of a continuous heterogeneity distribution, the needed use of simulated maximum likelihood due to intractable integrals. Already when there are more than four variables, these approaches can run into problems.[2] To avoid overfitting,

[1] Alternatively, one could allow for the responses to the different variables to be independent, to avoid needing at least 2^K outcomes. However, this assumption of independence can be too strict.

[2] In Hensher et al. (2013) the problem of estimation time is explicitly stated in footnote 5: it took over 100 hours to estimate the parameters based on a dataset with 588 units, 16 observations per unit and 4 variables that were allowed to be ignored.

Tang et al. (2020) use a penalization framework. They propose a linear model where each unit-specific parameter comes from a univariate discrete distribution with multiple possible outcomes of which one outcome is set to zero. The parameters of the discrete distributions are estimated by optimizing a penalized objective function. The idea of their approach can also be used for nonlinear models, but, in practice, the use of multiple univariate discrete distributions is too limited to capture the possible rich forms of heterogeneous responses, for example correlations across the responses to different variables.

The few papers that use a Bayesian approach also have their limitations. They are limited in terms of the underlying parametric model: only techniques for heterogeneous variable selection in the context of a multivariate linear, a multinomial logit, and a binary probit model have been proposed. Furthermore, the form of cross-sectional heterogeneity including heterogeneous variable selection is limited in these papers. Kim et al. (2018) let the unit-specific parameters come from a categorical distribution that simultaneously incorporates variable selection and other heterogeneity. Kim et al. (2009) follow a similar approach but instead consider a categorical distribution with an 'infinite' number of outcomes using a Dirichlet process prior. As with standard heterogeneous response models with discrete heterogeneity, the main drawback of the approaches of Kim et al. (2018) and Kim et al. (2009) is that the number of outcomes of the categorical distribution that is necessary to capture all combinations of variable selection is exponential in the number of explanatory variables. In practice, it is hard to find that many components.

A more parsimonious approach is developed in Gilbride et al. (2006), who let each unit-specific parameter be either equal to zero *or* come from an underlying multivariate normal distribution. However, this single multivariate normal distribution can be insufficient to describe the complex forms of unit-specific responses. Furthermore, the Markov chain Monte Carlo (MCMC) sampler that Gilbride et al. (2006) propose for posterior results can be computationally heavy when there are many variables, as in each MCMC iteration a likelihood function with 2^K terms has to be computed. Moreover, the MCMC sampler uses the prior distribution as candidate for drawing the unit-specific parameters. In case the data is quite informative, this candidate will have low acceptance rates and the sampler will have poor mixing.

In this paper, we generalize and improve the approach of Gilbride et al. (2006), thereby contributing to the literature in three important ways: by (i) generalizing to nonlinear models, (ii) substantially increasing the flexibility in the cross-sectional heterogeneity, and (iii) developing an efficient Bayesian MCMC sampler that also works well for up to 50 or 100 explanatory variables. The increased flexibility is obtained by augmenting the heterogeneous variable selection with an infinite mixture of multivariate normals using a Dirichlet process (DP) prior

To be more precise, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. For the model's unit-specific parameters we take a two-group approach: each unit-specific parameter is either zero or comes from a DP mixture of multivariate normals. In case of a single unit-specific parameter, such a two-group approach is referred to as a spike-and-slab prior (Mitchell & Beauchamp, 1988) or as stochastic search variable selection (SSVS) (George and McCulloch, 1993, George and McCulloch, 1997). We develop our approach for general nonlinear panel data models, encompassing multinomial logit and probit models, poisson and negative binomial count models, exponential models, among many others. The model is particularly useful in large N, small T settings, but can also be incorporated in large T settings because of the flexibility of the DP mixture.

We illustrate our approach with a Monte Carlo study and an empirical application. For illustration, we consider a multinomial logit model (MNL) as this model is the focus of most of the literature on heterogeneous variable selection. In the Monte Carlo study, we find that with our approach we can capture both complex forms of continuous cross-sectional heterogeneity — such as skewness and multimodality — as well as heterogeneous variable selection. When using only a 'standard' DP mixture for the unit-specific parameters, we find that heterogeneous variable selection cannot be accommodated. Instead of a spike at zero, this approach generally allocates substantial probability mass to parameter values in a relatively large interval around zero, depending on the shape of the true continuous heterogeneity distribution.

In the empirical application, we consider responses to a discrete choice experiment on food choices. We find substantial evidence of variable non-attendance and non-normality of

the continuous heterogeneity. In particular, the continuous heterogeneity distribution seems skewed. Hence, there seem to be quite some individuals that have strong preferences for certain attributes, and quite some individuals that ignore certain attributes. These findings indicate the usefulness of our approach in practice.

The setup of this paper is as follows. In Section 2, we discuss the related literature. In Section 3, we develop our approach for general nonlinear panel data models. We also provide the Bayesian MCMC sampler. In Sections 4 and 5, we discuss the results of our model for a small Monte Carlo study and an empirical application, respectively. In Section 6, we conclude.

2. Related literature

An overview of papers that develop approaches to accommodate heterogeneous variable selection in panel data models is given in Table 1. These papers mainly differ in (i) the type of model they develop (logit, probit, linear, et cetera), (ii) how they incorporate heterogeneous variable selection, (iii) how they deal with cross-sectional heterogeneity other than heterogeneous variable selection, and (iv) if and how they incorporate correlated variable selection.

Table 1: Overview of papers that develop approaches to accommodate heterogeneous variable selection.

Paper	Model	Het. variable selection	Additional cross-sectional heterogeneity	Correlated selection*
Frequentist				
Scarpa et al. (2009)	MNL	Latent class	Constant	Partly correlated
Hensher & Green (2010)	MNL	Latent class	Categorical	Partly correlated
Hole (2011)	MNL	Latent class	Constant	Partly correlated
Campbell et al. (2011)	MNL	Latent class	Categorical	Partly correlated
Hess et al. (2013)	MNL	Latent class	Multivariate normal	Uncorrelated
Hole et al. (2013)	MNL	Latent class	Multivariate normal	Partly correlated
Collins et al. (2013)	MNL	Latent class	Multivariate normal	Partly correlated
Hensher et al. (2013)	MNL	Latent class	Multivariate normal per latent class	Fully correlated
Tang et al. (2020)	Linear	Penalty	Categorical per variable	Uncorrelated
Bayesian				
Gilbride et al. (2006)	MNL	SSVS	Multivariate normal	Uncorrelated
Kim et al. (2009)	Linear	Spike-and-slab**	Categorical (infinite # of outcomes)	Uncorrelated
Kim et al. (2018)	Probit	Spike-and-slab**	Categorical	Uncorrelated
This paper	General	SSVS	Infinite mixture of multivariate normals	Uncorrelated

* The partly correlated methods are based on either considering only a subset of variables to be ignored together or letting the membership probabilities being a function of unit-specific variables.
** In Kim et al. (2009) and Kim et al. (2018), the underlying distribution for the unit-specific parameters incorporates heterogeneous variable selection within the categorical distribution that governs other cross-sectional heterogeneity.

Heterogeneous variable selection is mostly incorporated using a two-group approach (SSVS, spike-and-slab, latent class). The frequentist approaches rely on latent class tech-

niques (or a categorical distribution) for the unit-specific parameters. That is, these approaches specify 2^K classes where in each class a different combination of variables is selected, i.e. a different combination of parameters are set to zero. Each unit belongs to one of the 2^K classes. For the unit-specific parameters that are not zero, the approaches either restrict them to be equal over units (*constant*), allow them to differ depending on the class the unit is in (*categorical*), or let them be independent of the class a unit is in and let them come from an underlying multivariate normal distribution. Exceptions are Campbell et al. (2011) who use a single multivariate normal and additionally allow for a different scale parameter per class, and Hensher et al. (2013) who allow for a different multivariate normal per class. The Bayesian approaches rely on a spike-and-slab prior or stochastic search variable selection (SSVS). That is, when a variable is ignored/unselected, the corresponding unit-specific parameter is either zero (spike-and-slab prior) or comes from a distribution closely centered around zero (SSVS). Within the Bayesian approaches, Kim et al. (2009) and Kim et al. (2018) incorporate heterogeneous variable selection within the categorical distribution that describes other cross-sectional heterogeneity. In contrast, Gilbride et al. (2006) let these two types of heterogeneous responses be independent: a unit-specific parameter is either zero or comes from a separate multivariate normal distribution. Our approach is most similar to Gilbride et al. (2006). We extend upon their approach by generalizing to nonlinear models and using a Dirichlet process mixture of multivariate normals for the other heterogeneity to realistically capture differences across units. Moreover, we improve upon their MCMC sampler to allow the approach to be used for up to 50 or 100 explanatory variables.

Alternatively to the two-group approach, Tang et al. (2020) use a penalization framework to shrink the unit-specific parameters towards zero or towards a specific value out of a set of outcomes to be estimated. Similar penalization frameworks for heterogeneous variable selection are employed in image and video classification problems, see e.g. Wu et al. (2012) and Zhao et al. (2015), where the used term is often *heterogeneous feature selection* or *sparsification*. In contrast to the approach developed in Tang et al. (2020), these latter approaches shrink the corresponding unit-specific parameter to zero in case a variable is selected, and not to some underlying population distribution shared across units.

Another main difference between the available approaches for heterogeneous variable

selection is if and how they deal with correlated variable selection. Correlated variable selection refers to the phenomenon that some variables may be more likely to be selected/ignored together. This correlation can be divided into explained correlation (using observed unit-specific variables) and unexplained correlation. Most of the papers on heterogeneous variable selection do not allow for correlated variable selection. The ones that do can be divided into three groups: (i) letting each class/component have its own membership probability causing the number of membership probability parameters to be exponential in the number of explanatory variables (Hensher et al., 2013), (ii) allowing for variable selection and correlation only across predefined subsets of variables (Scarpa et al., 2009, Hensher and Greene, 2010, Campbell et al., 2011 and Collins et al., 2013), or (iii) letting the class membership probabilities be a function of unit-specific variables (Hole, 2011, Hole et al., 2013). In this paper, we do not explicitly allow for correlated variable selection. However, our approach can be extended to allow for both explained and unexplained correlated variable selection.

Approaches have also been developed that use a DP mixture for cross-sectional heterogeneity, and *aggregate* variable selection to analyze which variables should not be in the model for all units (see e.g. Cai and Dunson, 2005 and Yang, 2012). Furthermore, related approaches have been developed for models that do not include unit-specific parameters: the combination of a DP mixture and variable selection are used for a set of pooled parameters. These approaches are often used in settings with many explanatory variables to shrink coefficients towards zero (variable selection) or each other (DP mixture), both in supervised problems (see e.g. Dunson et al., 2008, MacLehose et al., 2007, and Korobilis, 2013) and unsupervised clustering problems (see e.g. Kim et al., 2006, Wang and Blei, 2009, Yu et al., 2010, Fan and Bouguila, 2013).

3. Methodology

In this section, we develop our approach to simultaneously allow for heterogeneous variable selection and other flexible forms of cross-sectional heterogeneity in nonlinear panel data models. We provide the model specification and the details of the MCMC sampler to obtain posterior samples.

We consider a dataset with N cross-sectional units and T_i observations for unit $i = 1, ..., N$. The interest is in modeling a scalar dependent random variable Y_{it} in terms of observed explanatory variables in x_{it} and z_{it} for unit i at time t. The responses to the variables in the $(K_x \times 1)$ vector x_{it} are assumed unit-specific and captured in the $(K_x \times 1)$ parameter vector β_i. For identification, x_{it} may contain time-varying variables only, other than an intercept.[3] The responses to the variables in the $(K_z \times 1)$ vector z_{it} are assumed equal across units and captured in the $(K_z \times 1)$ parameter vector γ. The variables in x_{it} and z_{it} cannot overlap.

We consider a nonlinear model for Y_{it} as given by

$$Y_{it}|\beta_i, \gamma \sim f(g(x_{it}, \beta_i, z_{it}, \gamma)), \tag{1}$$

where f is a known continuous or discrete probability distribution, g is a known (possibly multivariate) deterministic link function that maps x_{it}, β_i, z_{it} and γ to the parameters of the probability distribution, and we assume the observations Y_{it} to be conditionally independent over units and time periods.

For example, for multinomial data such as discrete choices, f could represent a multinomial distribution with size 1 and probability vector $p_{it} = g(x_{it}, \beta_i, z_{it}, \gamma)$ based on e.g. the softmax link function to obtain a multinomial logit model. For count data, f could represent a Poisson or negative binomial distribution with parameters $g(x_{it}, \beta_i, z_{it}, \gamma)$. Continuous distributions may also be used, such as the normal or the exponential distribution. We take the distribution $f()$ and the link function $g()$ as given.

The parameters in β_i capture the responses of unit i to the variables in x_{it}. To allow for flexible forms of cross-sectional heterogeneity, we take

$$\beta_{ik} = \tau_{ik}\lambda_{ik}, \tag{2}$$

for $k = 1, ..., K_x$. Heterogeneous variable selection is captured in the latent indicator τ_{ik}

[3] We recommend to mean center any continuous variable in x_{it}. Furthermore, for multinomial models, instead of a single intercept, x_{it} may contain an intercept per possible outcome for Y_{it}, minus one, or other time-invariant alternative-specific variables.

which indicates whether variable k is selected by unit i and, if selected, lets β_{ik} be equal to λ_{ik} which follows an infinite mixture of multivariate normals distribution using a Dirichlet process prior. We take $\tau_{ik} \in \{\kappa, 1\}$, where κ is zero or close to zero and is set by the researcher. In case $\kappa = 0$, we obtain a spike-and-slab prior, in case $\kappa \neq 0$ but close to zero our approach becomes an example of stochastic search variable selection. For estimation efficiency, it is not necessary to set $\kappa \neq 0$. Hence, for interpretation it may be most suitable to set $\kappa = 0$.

We assume the variable selection indicator (τ_{ik}) to be independent of λ_{ik}. The probability that unit i selects variable k is denoted by

$$\Pr[\tau_{ik} = 1|\theta_k] = \theta_k, \tag{3}$$

with $0 \leq \theta_k \leq 1$, for $k = 1, ..., K_x$.[4]

For flexible continuous heterogeneity, we let $\lambda_i = (\lambda_{i1}, ..., \lambda_{iK_x})'$ come from an infinite mixture of multivariate normals using the DP prior (Ferguson et al., 1974, Antoniak, 1974, Rossi, 2014). The mixture for λ_i is given by

$$\lambda_i|\{\pi_q\}_q, \{\mu_q\}_q, \{\Sigma_q\}_q \sim \sum_{q=1}^{\infty} \pi_q MVN(\mu_q, \Sigma_q), \tag{4}$$

where π_q indicates the component membership probability of component q, μ_q denotes component's q mean, and Σ_q denotes component's q covariance matrix. The DP prior puts a prior on the mixture parameters π_q, μ_q and Σ_q. The DP prior has two hyperparameters: a tightness parameter α and a base distribution G_0 that invoke the following priors on π_q, μ_q and Σ_q

$$\pi_q = \eta_q \prod_{r=1}^{q-1}(1 - \eta_r), \quad \eta_q \sim Beta(1, \alpha), \tag{5}$$

$$\mu_q, \Sigma_q \sim G_0 \equiv p(\mu_q, \Sigma_q), \tag{6}$$

[4]One can allow for explained correlated variable selection using unit-specific probabilities θ_{ik} that are a deterministic function of unit-specific variables.

for $q = 1, 2, ...$, where the base distribution G_0 of the DP is the prior distribution $p(\mu_q, \Sigma_q)$. This representation of the DP mixture is known as the stick-breaking representation (Rossi, 2014).

The component membership probabilities π_q are completely governed by the tightness parameter α. The specification implies that π_q declines as the component indicator q increases. The larger α, the more mass the Beta distribution has at zero. Hence, the larger α, the smaller we expect the η_q's for the first components to be, and the more components we expect to have reasonably large membership probabilities. Given that there are N units, at most N unique components can be identified from the data.

For the base distribution, we take the conjugate prior $p(\mu_q, \Sigma_q) = p(\mu_q|\Sigma_q)p(\Sigma_q)$ as given by

$$p(\mu_q|\Sigma_q) = MVN(\mu_0, d^{-1}\Sigma_q), \tag{7}$$

$$p(\Sigma_q) = IW(\nu, \nu\upsilon I). \tag{8}$$

This conjugate prior allows for efficient estimation. The hyperparameter υ affects the variances of the components: a large υ puts substantial prior mass on components with 'large' variance, whereas a small υ puts substantial prior mass on components with 'small' variance (Rossi, 2014).

Finally, for γ and θ_k we take the following priors

$$p(\gamma) = MVN(\gamma_0, \Sigma_\gamma), \tag{9}$$

$$p(\theta_k) = Beta(a, b), \qquad \text{for } k = 1, ..., K_x. \tag{10}$$

The hyperparameters α, μ_0, d, υ, ν, γ_0, Σ_γ, a and b should either be set by the researcher or should have a prior itself. The proposed approach for heterogeneous responses is particularly useful in large N, small T settings, but can also be incorporated in large T settings because of the flexibility of the DP mixture.

As a final remark, we note that one may wish to restrict the variable selection to hold for multiple variables simultaneously. For example, in case one includes different levels of

the same categorical variable through multiple dummy variables, one may want the variable selection to hold for all levels of that categorical variable. More formally, some of the elements in $\tau_i = (\tau_{i1}, ..., \tau_{iK_x})'$ should be allowed to be restricted to be equal to one another. Such restrictions can be incorporated by introducing the unknown $(K_x^* \times 1)$ vector τ_i^* with elements that can all differ from each other, and a known $(K_x \times K_x^*)$ selection matrix D^* to correctly map τ_i^* to τ_i via $\tau_i = D^* \tau_i^*$, where $K_x^* \leq K_x$. The selection matrix D^* should be set by the researcher, its elements are either zero or one, and it can have only a single one per row. In case $D^* = I_{K_x}$ we obtain the original formulation. Details of the prior specification and inference can be easily adapted.

3.1. Inference

For inference, we develop an efficient Bayesian MCMC sampler. The details of the MCMC sampler are outlined in Appendix A. Specialized code was written in R and C++ to obtain the posterior samples.[5] In this section, we present the main ideas.

To draw the DP mixture parameters, we use algorithm 2 in Neal (2000). That is, we augment the parameter space with the latent membership indicator c_i that indicates which mixture component unit i belongs to. This procedure is similar to that for a finite mixture, except that for the DP mixture, components may appear or disappear in subsequent MCMC iterations. Due to the conjugacy of the base distribution $p(\mu_q, \Sigma_q)$, we can use a computationally efficient Gibbs step to draw c_i. Moreover, in this Gibbs step we draw c_i unconditional on the component membership probabilities π. Hence, there is no need to draw π.

[5]The code for the MCMC sampler was tested using the identity (Geweke, 2004 and Cook et al., 2006)

$$p(\omega) = \int p(\omega|\tilde{y}) p(\tilde{y}|\tilde{\omega}) p(\tilde{\omega}) d\tilde{y} d\tilde{\omega}$$

where ω are the model parameters, $\tilde{\omega}$ is a draw from the prior density $p(\omega)$, $\tilde{y}$ is a draw from the DGP with likelihood function $p(y|\tilde{\omega})$ given $\tilde{\omega}$, and $p(\omega|\tilde{y})$ is the posterior density of ω given $\tilde{y}$. During testing, we used many replications to approximate the integral on the right-hand side and checked whether the approximated marginal densities of ω matched the prior marginal densities. That is, for each replication, we drew $\tilde{\omega}$ from its prior and used this draw to generate data $\tilde{y}$ from the DGP. Next, we used the MCMC sampler to obtain posterior draws for ω given the generated data $\tilde{y}$. Finally, for each parameter in ω, we considered the posterior draws over all replications, and checked whether the posterior marginal densities coincided with the prior marginal densities.

Per MCMC iteration, we draw (i) the DP mixture parameters $\{\lambda_i\}_{i=1}^N$, $\{c_i\}_{i=1}^N$, $\{\mu_q\}_q$ and $\{\Sigma_q\}_q$, (ii) the variable selection parameters $\{\tau_i\}_{i=1}^N$ and θ, and (iii) γ. Conditional on $\{c_i\}_{i=1}^N$, drawing $\{\lambda_i\}_{i=1}^N$, $\{\mu_q\}_q$ and $\{\Sigma_q\}_q$ becomes straightforward: λ_i can be drawn using a random walk Metropolis-Hastings (M-H) step (Metropolis et al., 1953, Hastings, 1970), μ_q can be drawn from a multivariate normal using only the λ_i from the units for which $c_i = q$, and similarly Σ_q can be drawn from an inverse Wishart distribution. Furthermore, we draw γ using a random walk M-H step, τ_{ik} using a Bernoulli distribution, and θ_k from a Beta distribution.

For some models, including the linear model, the M-H steps to draw λ_i and γ can be directly replaced by Gibbs steps. For models in which this is not the case, we do not recommend to perform any further data augmentation to enable a Gibbs step for λ_i and γ. For example, we would not recommend to augment the latent utilities in the multinomial logit model (using e.g. the augmentation schemes in Polson et al., 2013 or Frühwirth-Schnatter and Frühwirth, 2010). Such types of data augmentation can lead to poor mixing in the MCMC sampler. The main reason for poor mixing is that, for the example of the multinomial logit model, the latent utilities are drawn conditional on the variable selection indicators τ_i. In case in a MCMC iteration, one obtains a draw $\tau_{ik} = 0$, the draw for the latent utility will assign no weight to the k^{th} variable. In the next MCMC iteration, this may cause a high probability to again draw $\tau_{ik} = 0$ conditional on the latent utility. That is, the correlation between posterior draws of τ_i and the latent utilities can be quite high.

To improve mixing of the sampler, we jointly draw λ_{ik} and τ_{ik} for each variable k, and we randomize the order over k across the MCMC iterations. Alternatively, one may jointly draw λ_i and τ_i over all variables. In that case, the computation of the likelihood function requires the evaluation of 2^{K_x} terms of likelihood contributions of unit i due to all possible combinations of variables selected. These evaluations can generally not be simplified. Hence, this should only be done when K_x is small, say smaller than five. By drawing separately per variable, the likelihood function contains only 2 terms to compute (one for $\tau_{ik} = 1$ and one for $\tau_{ik} = \kappa$) and this has to be repeated K_x times.

Our model and Bayesian MCMC sampler can be used for any nonlinear model of the form in Equation (1). The sampler does rely on the computation of the likelihood function

conditional on λ_i and γ, for performing the M-H steps for λ_{ik} and γ and for drawing τ_{ik}. For many models, this likelihood function can be analytically computed, e.g. for the multinomial logit model, poisson model, and negative binomial model. For other models, the likelihood function has to be approximated, e.g. for the multinomial probit model (MNP) when the number of possible outcomes for Y_{it} exceeds two. For these later cases, our MCMC sampler can become slow due to the computations necessary for approximating the likelihood function, and more efficient approaches could entail further data augmentation, for example the latent utilities for the MNP. Again, care must be taken, because conditioning on the augmented parameters can lead to high correlation in the chains due to the conditioning on the variable selection indicators τ_i.

4. Monte Carlo study

In this section, we perform a small Monte Carlo study to examine the performance of our proposed approach for accommodating heterogeneous variable selection. For this purpose, we consider a multinomial logit model (McFadden, 1973, Manski, 1977). At each observation t, a unit i selects one of J alternatives. Each alternative j is described by K_x variables in the vector x_{itj}. The multinomial logit model is given by

$$Y_{it} \sim \text{Multinomial}(1, p_{it}), \tag{11}$$

$$p_{itj} \equiv \Pr[Y_{it} = j|\beta_i] = \frac{\exp(x'_{itj}\beta_i)}{\sum_{l=1}^{J}\exp(x'_{itl}\beta_i)}, \qquad j = 1, ..., J, \tag{12}$$

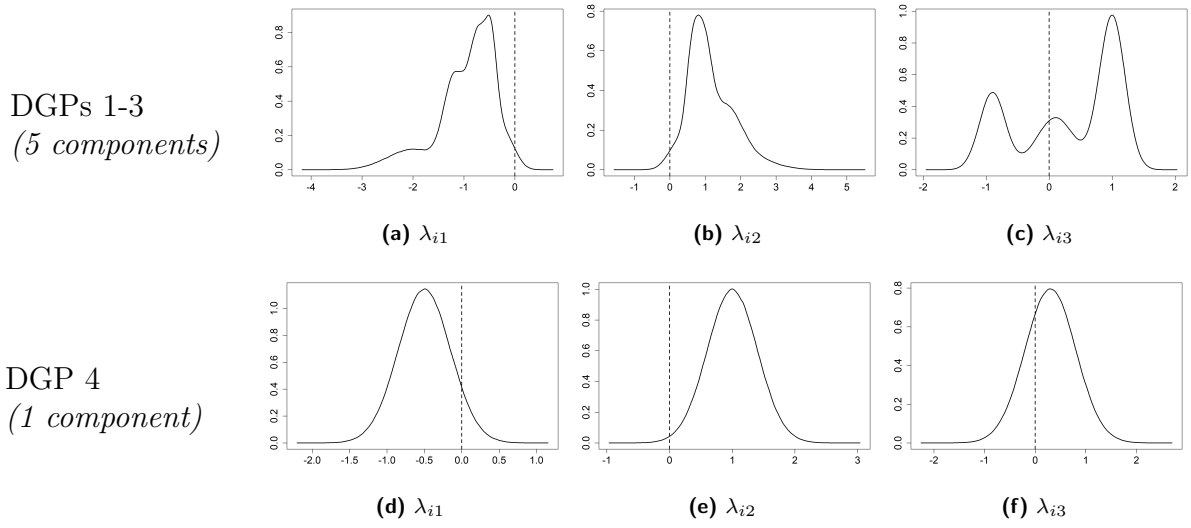
where $p_{it} = (p_{it1}, ..., p_{itJ})'$.

We consider four data generating processes (DGPs) and perform 100 Monte Carlo replications per DGP. In each DGP, we consider $1,000$ units, 20 observations per unit, 3 alternatives per observation, and 3 variables: x_{1itj} from a standard normal distribution and x_{2itj}, x_{3itj} from a Bernoulli distribution with probability of outcome 1 equal to 0.5. For all DGPs, we let $\beta_{ik} = \tau_{ik}\lambda_{ik}$, where $\tau_{ik} \in \{0, 1\}$ is the variable selection indicator, for $k = 1, 2, 3$.

For DGPs 1 to 3, we let λ_i come from a mixture of multivariate normals with five components. The components' means, covariance matrices and weights are equal across

the three DGPs, whereas the amount of variable selection differs across the DGPs. In the mixture, the marginal density of λ_{i1} mostly has mass on the negative domain, is skewed and a has an extra mode in the tail, that of λ_{i2} is skewed with mass mostly on the positive domain, and that of λ_{i3} is multimodal with a mode at zero and substantial mass on both the positive and negative domain, see Figures 1 (a)-(c).[6] Hence, the first variable could represent price, the second variable a quality indicator, and the third variable a brand indicator. For the heterogeneous variable selection part, we take the following probabilities that a variable is relevant for a unit, i.e., that the unit assigns weight to the variable. In DGP 1, the variables are relevant for the majority of units: $\theta = (0.90, 0.85, 0.95)$. In other words, 90% of units assign weight to the first variable, 85% to the second variable, and 95% to the third variable. In DGP 2, the variables are relevant for all units: $\theta = (1, 1, 1)$. In DGP 3, there are quite some units for which the variables are irrelevant: $\theta = (0.80, 0.70, 0.75)$.

Figure 1: True marginal densities of λ_{i1}, λ_{i2} and λ_{i3} for DGPs 1 to 3 (top) and DGP 4 (bottom).



(a) λ_{i1} **(b)** λ_{i2} **(c)** λ_{i3}

(d) λ_{i1} **(e)** λ_{i2} **(f)** λ_{i3}

For DGP 4, we use one mixture component for λ_i, see Figures 1 (d)-(f).[7] We use the same amount of variable selection as in DGP 1, that is, $\theta = (0.90, 0.85, 0.95)$.

[6]For DGPs 1-3 with five mixture components we use the following setting. We set the membership probabilities to $\pi = (0.25, 0.1, 0.15, 0.1, 0.4)$, the components' means to $\mu_1 = (-1.2, -0.45, -2, -0.2, -0.7)$, $\mu_2 = (1.6, 0.6, 2, 0.25, 0.9)$ and $\mu_3 = (0.1, 1, -0.9, -0.9, 1)$, and the components' covariance matrices with standard deviations, $\sigma_1 = (0.2, 0.1, 0.5, 0.2, 0.2)$, $\sigma_2 = (0.4, 0.15, 0.75, 0.3, 0.25)$, and $\sigma_3 = (0.3, 0.2, 0.2, 0.2, 0.2)$, and correlations (equal across components) $\rho_{12} = 0.2$, $\rho_{13} = 0.1$ and $\rho_{23} = 0.2$.

[7]For DGP 4 with one mixture component we set the mean to $\mu = (-0.5, 1.0, 0.3)$ and the covariance matrix to Σ with standard deviations $\sigma = (0.35, 0.40, 0.50)$ for the three variables, respectively, and correlations $\rho_{12} = 0.2$, $\rho_{13} = 0.1$ and $\rho_{23} = 0.4$.

We estimate a MNL using three different approaches for the heterogeneous responses: (1) our proposed DP mixture with heterogeneous variable selection (HVS-DPM), (2) a "standard" DP mixture without heterogeneous variable selection (DPM), and (3) a *single* multivariate normal distribution with heterogeneous variable selection (HVS-M). We set the priors' hyperparameters to $\alpha = 1$, $\mu_0 = 0$, $d = 0.5$, $\nu = K_x + 5$, $\upsilon = 0.2$, and $a = b = 1$. Hence, the prior distribution for θ_k is uniform over the unit interval. Appendix B gives the histograms of the prior number of components based on α and N, the marginal prior on μ and the marginal prior on the standard deviations on the diagonal of Σ. Furthermore, we set $\kappa = 0$ in estimation.

For the posterior results per replication, we use 15,000 simulations after 5,000 burn-in draws and keep every 4^{th} draw. We visualize the results per DGP using the posterior marginal densities of β_{i1}, β_{i2}, and β_{i3}. For this purpose, we first construct the posterior marginal densities for each of the 100 replications. That is, for each replication, we take the equally weighted mixture of the 15,000/4 posterior draws of marginal densities, where each draw of the density directly results from the draws of the parameters of the mixture of multivariate normals (π, μ, Σ) and of the heterogeneous variable selection (θ). For each DGP, we plot the equally weighted mixture of these 100 marginal densities.

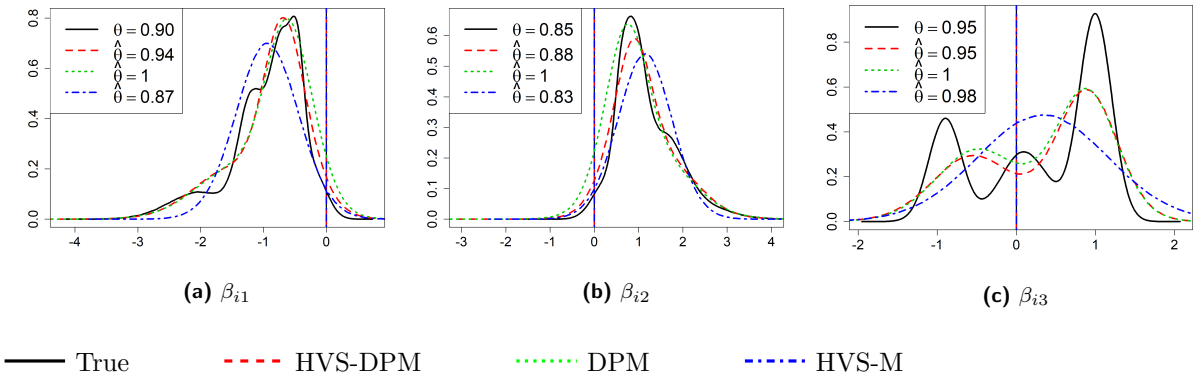
4.1. Results

The posterior results for DGP 1, with substantial variable relevance and non-normal continuous heterogeneity, are shown in Figure 2.[8] In this figure, we plot the marginal posterior densities of β_{i1}, β_{i2}, and β_{i3}, by plotting the underlying continuous heterogeneity distribution (the mixture of multivariate normals) as a continuous density. Moreover, we represent the heterogeneous variable selection, i.e. the relative number of units that assign no weight to the variable, by a vertical line through zero. The probability mass at zero is equal to one minus the mean across replications of the posterior mean of θ, displayed in the top left corner.

Our proposed model is well able to capture the skewness and multimodality in the contin-

[8]To obtain 1.000 draws from the posterior for a Monte Carlo replication generated from DGP 1, it takes about 50 seconds for the HVS-DPM, 10 seconds for the DPM and 45 seconds for the HVS-M. Simulations were done using 1 core on an Intel Core i7 processor with 2.6GHz frequency.

Figure 2: (Posterior) marginal densities of β_{i1}, β_{i2} and β_{i3} for DGP 1.



(a) β_{i1} **(b)** β_{i2} **(c)** β_{i3}

uous heterogeneity. The fit is not perfect, mainly because we find components that are less peaked than they are in reality, that is, we find components with larger variances. Due to this smoothing, primarily caused by the prior on the covariance matrices, the mass close to zero of the continuous heterogeneity distribution is slightly overestimated and therefore the probability that a variable is selected is overestimated. In sum, for the skewed distribution for variables one and two, our model is able to capture the modes and the heavy tails. For variable three, the mode at zero of the continuous heterogeneity distribution is missed, and the modes at the positive and negative side are less extreme than in reality.

Compared to the alternative approaches, our approach seems to best capture the underlying distribution of heterogeneous responses. The standard DP mixture without variable selection cannot capture the spike at zero. Instead, more mass is allocated between -0.5 and 0.5. The single multivariate normal approach with variable selection cannot capture the non-normality in the continuous heterogeneity, and compensates by shifting the mode away from zero for the skewed distributions, and finding much less heavy tails.

To further compare the performance of the three approaches for modeling heterogeneous responses, we consider the predictive log-likelihood. We generate five more observations for each unit. For each Monte Carlo replication and each approach, we compute the predictive log-likelihood based on these five out-of-sample observations per unit.[9] For easy comparison,

[9]The predictive log-likelihood is computed using the posterior samples:

$$\sum_{i=1}^{N} \log \left[\frac{1}{S} \sum_{s=1}^{S} \prod_{t=21}^{25} \Pr[Y_{it} = y_{it} | \beta_i^{(s)}] \right], \tag{13}$$

we subtract the log-likelihood value obtained with one of the alternative approaches (DPM or HVS-M) from the value obtained with our approach (HVS-DPM). A positive number indicates our approach leads to a better predictive performance, a negative number indicates the alternative approach leads to a better predictive performance.

Table 2: Difference between the predictive log-likelihood of the MNL using the HVS-DPM for heterogeneous responses against two alternative approaches for heterogeneous responses (DPM and HVS-M) per DGP. Based on 100 replications. Averages and percentages of replications for which difference is greater than zero.

	HVS-DPM against DPM		HVS-DPM against HVS-M	
DGP	Mean	% > 0	Mean	% > 0
DGP 1	2.3	84%	24.9	100%
DGP 2	-0.5	41%	37.0	100%
DGP 3	5.5	98%	11.4	98%
DGP 4	1.0	75%	-0.2	36%

The results on the predictive performance are in Table 2. We report the means over the Monte Carlo replications and the fraction of Monte Carlo replications for which our approach has a better predictive performance according to the predictive log-likelihood. For DGP 1, we find that the predictions obtained with our approach are substantially better than those obtained with the alternative approaches. This holds in particular in comparison with the single multivariate normal approach (HVS-M): none of the replications of the HVS-M approach has a higher log-likelihood value.

For further evaluation, we consider the hit rates: how well are the MNLs based on the three approaches able to accurately assign, at the unit-level, posterior mass to β_{ik}. The results are in Table 8. In this table, we show the percentage of units for which the posterior draw of β_{ik} lies in the interval $[-\epsilon, \epsilon]$ for different values of ϵ, averaged over draws, variables and replications. We do this for four groups: (1) all units, (2) units for which the true β_{ik} lies within the interval, (3) units for which the true β_{ik} does not lie within the interval, and (4) units for which the true $\beta_{ik} = 0$. For DGP 1, we find that our approach slightly underestimates the mass between [-0.3,0.3], but not as much as the standard DP mixture approach. In contrast, the single multivariate normal approach leads to an overestimation

where S is the number of draws of the MCMC sampler after burn-in and $\beta_i^{(s)}$ is the s^{th} posterior draw of β_i which can be computed directly using the s^{th} posterior draws for δ_i and τ_i.

of the mass close to zero, and underestimation of the mass in the tails. Because of this, the HVS-M approach is better able to assign posterior mass to units that assign weights close to zero but does worse for units with weights further away from zero.

Table 3: Percentage of units for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ for multiple values of ϵ (averaged over Monte Carlo replications, draws and variables) for DGP 1. The results for DGPs 2 to 4 are in Appendix C.

	(1) All				(2) True $-\epsilon \leq \beta_{ik} \leq \epsilon$			(3) True $\beta_{ik} < -\epsilon$ or $\beta_{ik} > \epsilon$			(4) True $\beta_{ik} = 0$		
ϵ	True	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M
0.00	10	7	0	10	24	0	36	6	0	8	24	0	36
0.10	13	11	5	15	29	13	38	8	4	11	31	12	42
0.20	17	14	10	19	36	25	44	10	7	14	39	25	48
0.30	20	19	16	23	43	36	49	13	11	17	46	36	54
0.40	23	24	22	28	49	45	55	16	15	20	54	47	60
0.50	28	29	29	33	55	53	59	19	19	24	61	57	66
0.75	43	46	47	48	67	68	68	30	32	32	77	76	78
1.00	64	64	65	62	78	79	76	40	41	38	88	88	87
1.50	88	86	86	85	92	92	90	42	44	52	97	97	97
2.00	95	94	94	96	96	96	97	59	60	81	99	99	99
2.50	98	98	98	99	98	98	100	78	77	95	100	100	100

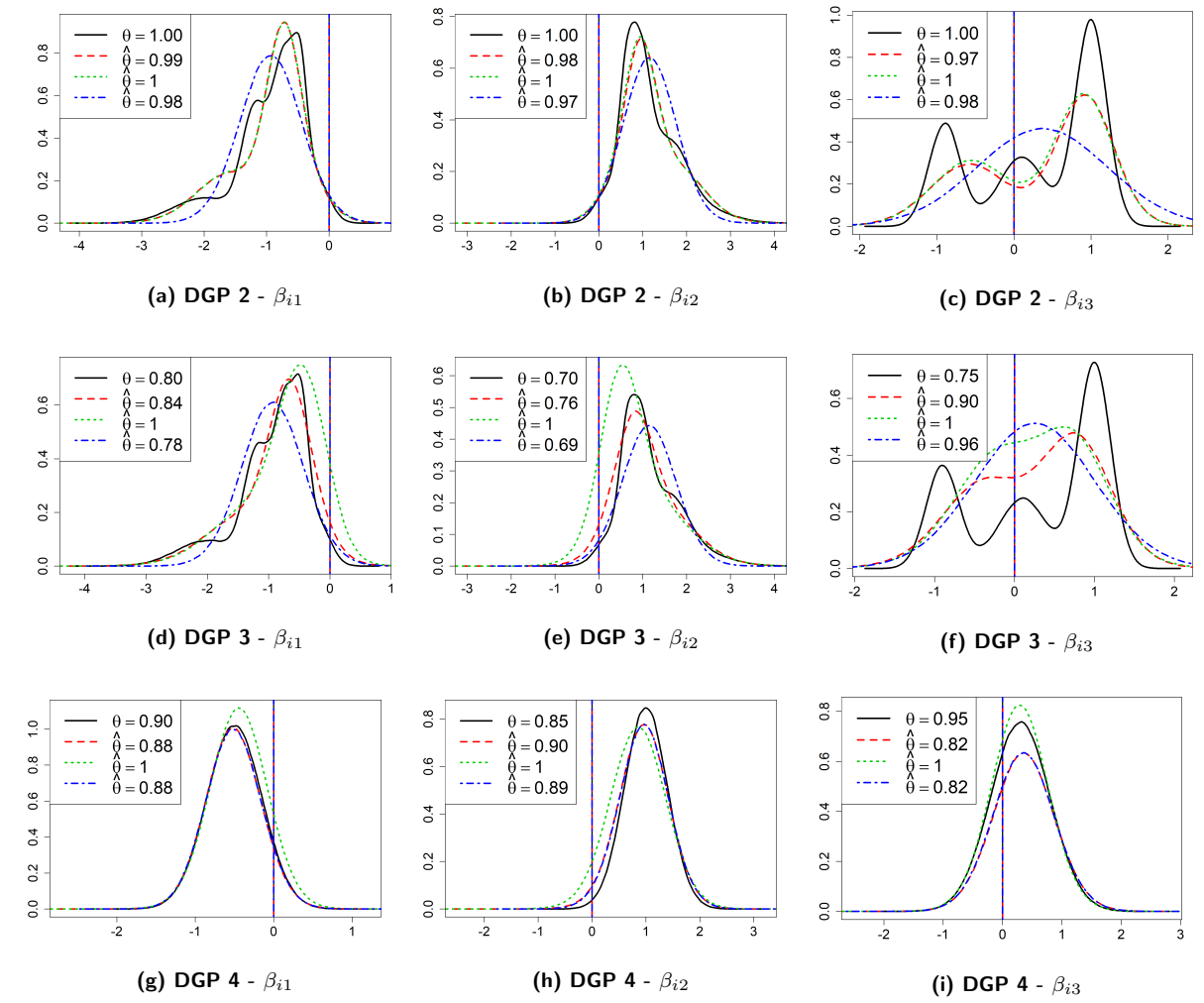
Results for four different groups:
(1) all units,
(2) units for which true β_{ik} falls within interval,
(3) units for which true β_{ik} does not fall within interval,
(4) units for which true $\beta_{ik} = 0$ ($\tau_{ik} = 0$).

For DGP 2 where the variables are relevant for all units, we find that, as expected, the results of our approach closely match those of the standard DP mixture approach, see Figures 3 (a)-(c). With our approach, we do find evidence of a small amount of units which do not assign weight to certain variables (1%-3%). The predictive log-likelihoods indicate that our approach leads to a similar predictive performance as the standard DP mixture, with the standard DPM being slightly better, see Table 2.

The results for DGP 3, where for quite some units the variables are irrelevant, are in Figures 3 (d)-(f). Again, we find that our approach seems to be most accurate in capturing the true density. Furthermore, the improvement in predictive performance of our approach as compared to the standard DP mixture approach is greater than for DGP 1. Hence, the more units that assign no weight to certain variables, the more important it becomes to account for heterogeneous variable selection.

When the continuous heterogeneity follows a normal distribution as in DGP 4, our approach and the HVS-M approach with a single multivariate normal for the continuous heterogeneity find a similar shape for the underlying distribution of heterogeneous responses, see Figures 3 (g)-(i). For the third variable, the amount of variable selection is underesti-

Figure 3: (Posterior) marginal densities of β_{i1}, β_{i2} and β_{i3} for DGPs 2-4.



(a) DGP 2 - β_{i1} **(b) DGP 2 - β_{i2}** **(c) DGP 2 - β_{i3}**

(d) DGP 3 - β_{i1} **(e) DGP 3 - β_{i2}** **(f) DGP 3 - β_{i3}**

(g) DGP 4 - β_{i1} **(h) DGP 4 - β_{i2}** **(i) DGP 4 - β_{i3}**

True
HVS-DPM
DPM
HVS-M

mated by both approaches: an estimated 82% of units assign weight to the third variable, whereas in reality it is 95%. This affects the shape of continuous heterogeneity found, which underestimates the mass between -0.5 and 0.5. As expected, the predictive log-likelihoods in Table 2 indicate that our approach leads to a similar predictive performance as the HVS-M approach.

As a final note. In this Monte Carlo study, we use $K_x = 3$ variables. Already with this small number of variables, we see that our approach with heterogeneous variable selection performs better than the standard DP mixture approach. In case there are more variables, we expect this difference in performance to be even greater, as the standard DP mixture would need at least 2^{K_x} components to capture all combinations of variable selection.

5. Case study: multinomial logit model

In this section, we illustrate our approach with an empirical application. We again consider the multinomial logit model in Equations (11) and (12). We consider responses obtained from a discrete choice experiment on food choices (Koç & van Kippersluis, 2017).[10] During the choice experiment, respondents had to complete 18 choice tasks. In each task, a respondent was asked which out of two meals s/he would eat most regularly. The meals were described by attributes as price and taste, and by attributes describing how healthy the meal is.

The respondents were divided into three groups. Each respondent group obtained different types of choice tasks in terms of the attributes describing how healthy the meal is and the amount of health information provided in the text. For group 1 (1,206 respondents), the meals were described by four attributes: price, cooking time, taste, and health consequences. All health information was provided in the final attribute health consequences. For groups 2 (1,154 respondents) and 3 (1,185 respondents), the meals were described by six attributes: price, cooking time, taste, number of calories, grams of saturated fat, and grams of sodium. Group 2 obtained health information in the text regarding what amount of calories, saturated fat, and sodium constitutes a healthy meal, whereas group 3 did not obtain health information. The ordering of the tasks within each respondent group was randomized over

[10]We thank the LISS panel and the experiment designers for providing this dataset.

the respondents.

Table 4: Attributes and attribute levels in the choice tasks for the discrete choice experiment on healthy food choices. The final column indicates which respondents groups (1,2 or 3) saw which attributes in the choice experiment.

Attribute	Attribute levels			Respondent groups
Price	2 Euro	6 Euro	10 Euro	1, 2, 3
Cooking time	10 min	30 min	50 min	1, 2, 3
Taste	OK	Good	Very good	1, 2, 3
Health consequences	Unhealthy	Health neutral	Healthy	1
Number of kilocalories	800	1,100	1,400	2, 3
Grams of saturated fat	10	20	30	2, 3
Milligrams of sodium	900	1,200	1,500	2, 3

Each of the attributes took on one of three values. The attribute levels had a clear ordering, see Table 4. For example, the price of the meal could either be 2 Euros, 6 Euros, or 10 Euros. In the model, we include a separate dummy variable per attribute level, with the exception of a baseline level per attribute (the middle level). Furthermore, we restrict the variable selection to hold for all levels of the same attribute. That is, we consider whether an individual finds an attribute relevant (such as price), and not just one of the attribute levels (such as price 2 Euros). Heterogeneous variable selection in such an application is also known as attribute non-attendance (Scarpa et al., 2009).

As in the Monte Carlo study, we use three approaches for modeling heterogeneous responses in the MNL: (1) our proposed DP mixture with heterogeneous variable selection (HVS-DPM), (2) a "standard" DP mixture without heterogeneous variable selection (DPM), and (3) a single multivariate normal distribution with heterogeneous variable selection (HVS-M). For posterior results, we use 60,000 simulations after 40,000 burn-in draws and we keep every 10^{th} draw. We use the same priors as in the Monte Carlo study and set $\kappa = 0$.

The MCMC sampler converges rather quickly and mixes well in general. For extreme quantiles of the heterogeneity distribution, the mixing is less good. This is not surprising as only very few observations are informative for such quantiles. Trace plots are given in the Supplementary Materials, available upon request.

5.1. Results

The posterior marginal densities of β_i for the first respondent group are displayed in Figure 4.[11] For this group, the meals were described by four attributes. Using our approach with a DP mixture and heterogeneous variable selection, we find evidence of the existence of groups of respondents that 'ignore' attributes, for all four attributes. Ignorance of attributes, or attribute non-attendance, can mean that either a respondent did not consider the attribute or is indifferent between the attribute levels. The health attribute is least ignored (4%), followed by price (13%), taste (16%), and cooking time (33%). The marginal distributions seem skewed, most mass is usually at either the positive or the negative side, and there is a heavy tail away from zero. For the health attribute levels, the tail is especially thick, indicating that there are groups of respondents that highly value this attribute.

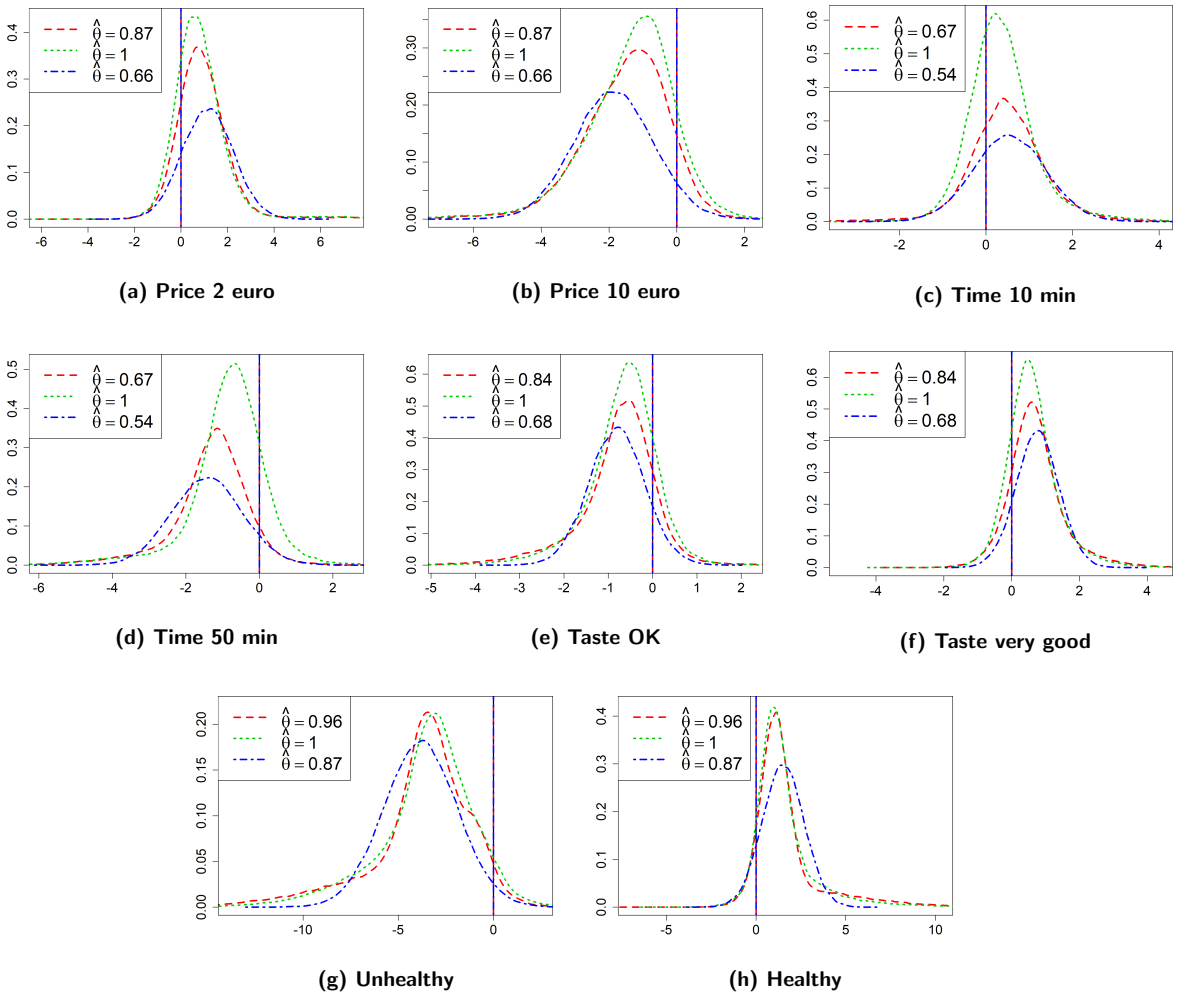
The HVS-M approach with a single multivariate normal clearly cannot capture the skewness in the marginal distributions. Instead, to somewhat capture the heavy tail and that most mass is on one side of the distribution, the mode of the distribution is shifted further away from zero, leading to selection probabilities that are substantially lower than we find with our approach. Finally, the standard DP mixture without variable selection finds roughly the same forms of the density as our approach with variable selection, but as it cannot capture the peak at zero, it distributes more mass between -1.0 and 1.0. This can be seen most clearly in Table 5, which shows the percentage of draws for β_{ik} in the interval $[-\epsilon, \epsilon]$. In this table, we also see that the DP mixture approaches assign more mass in the tails than the HVS-M approach.

The results for the second (health info) and third (no health info) respondent groups are in Figures 5 and 6, respectively. For these groups, the meals were described by six attributes. For both respondent groups we again find evidence of variable ignorance and non-normality of the heterogeneity when using our approach.

To better show the difference in variable selection across the respondent groups, Table 6 concisely displays the posterior means and 95% highest posterior density intervals (HPDIs) of θ_k — the probability that attribute k is selected — per respondent group and attribute for the

[11] The posterior marginal densities are constructed from the posterior draws in the same way as in the Monte Carlo study.

Figure 4: Posterior marginal density of β_{ik} for respondent group 1.



(a) Price 2 euro **(b) Price 10 euro** **(c) Time 10 min**

(d) Time 50 min **(e) Taste OK** **(f) Taste very good**

(g) Unhealthy **(h) Healthy**

Baseline levels are price 6 euro, cooking time 30 minutes, taste good, and health neutral.

- - - - HVS-DPM

····· DPM

-·-·- HVS-M

Table 5: Percentage of individuals for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ (averaged over draws and variables) per respondent group.

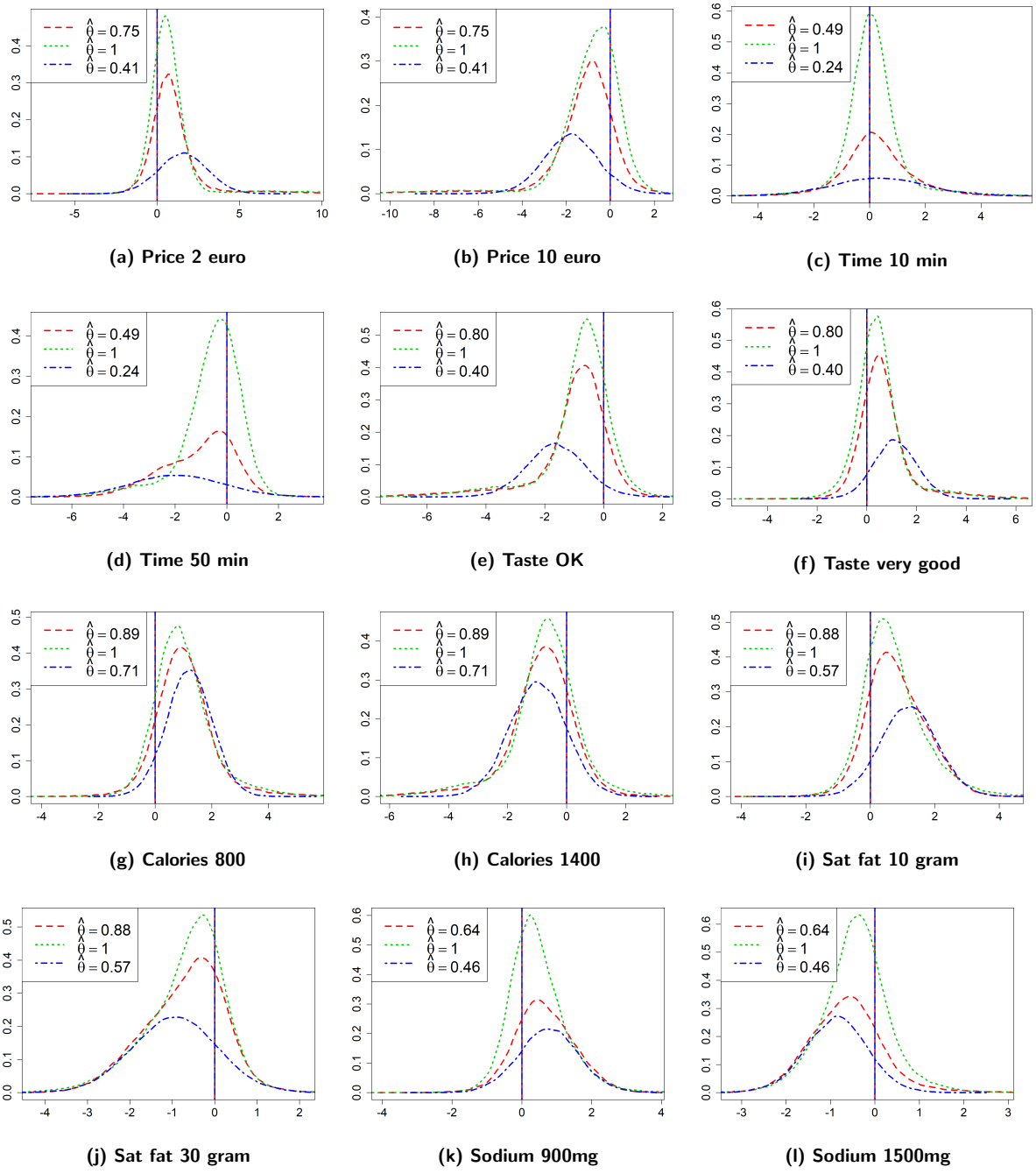
	Group 1			Group 2			Group 3		
ϵ	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M
0.00	17	0	31	26	0	54	33	0	58
0.10	20	6	34	31	9	55	37	9	59
0.20	24	12	36	36	17	57	42	18	61
0.30	28	18	39	41	25	59	46	27	63
0.40	32	24	42	46	33	61	50	35	64
0.50	36	30	45	50	41	63	55	43	66
0.75	46	43	52	60	57	68	64	59	71
1.00	55	54	58	70	69	73	72	71	75
1.50	70	70	70	83	84	83	84	85	83
2.00	79	79	79	90	91	90	90	91	89
2.50	84	85	85	94	95	95	93	93	94
3.00	88	89	90	96	96	97	95	95	97
4.00	93	93	94	98	98	99	97	97	99
5.00	96	96	97	99	99	100	99	99	100

DP mixture approach with heterogeneous variable selection. For the meals described by six attributes (groups 2 and 3), including the health information seems to have the respondents made more aware of calories and saturated fat, but the opposite seems to hold for sodium. Furthermore, compared to the first group, the individuals in the second and third group seem to more often ignore the standard attributes price, cooking time, and taste. The 95% HPDIs are quite wide, indicating that there is quite some uncertainty in these values.

Table 6: Posterior means and 95% HPDIs of attribute selection probabilities θ per respondent group and attribute (results of HVS-DPM).

	Group 1		Group 2		Group 3	
Attribute	Mean	95% HPDI	Mean	95% HPDI	Mean	95% HPDI
Price	0.87	(0.79,0.96)	0.75	(0.65,0.84)	0.63	(0.52,0.76)
Cooking time	0.67	(0.57,0.77)	0.49	(0.39,0.58)	0.39	(0.31,0.48)
Taste	0.84	(0.72,0.97)	0.80	(0.67,0.97)	0.75	(0.65,0.85)
Health	0.96	(0.93,0.99)	-		-	
Number of kilocalories	-		0.89	(0.80,0.98)	0.81	(0.72,0.89)
Grams of saturated fat	-		0.88	(0.78,1.00)	0.86	(0.73,0.97)
Milligrams of sodium	-		0.64	(0.51,0.79)	0.74	(0.56,0.91)

Figure 5: Posterior marginal density of β_{ik} for respondent group 2.



(a) Price 2 euro (b) Price 10 euro (c) Time 10 min

(d) Time 50 min (e) Taste OK (f) Taste very good

(g) Calories 800 (h) Calories 1400 (i) Sat fat 10 gram

(j) Sat fat 30 gram (k) Sodium 900mg (l) Sodium 1500mg

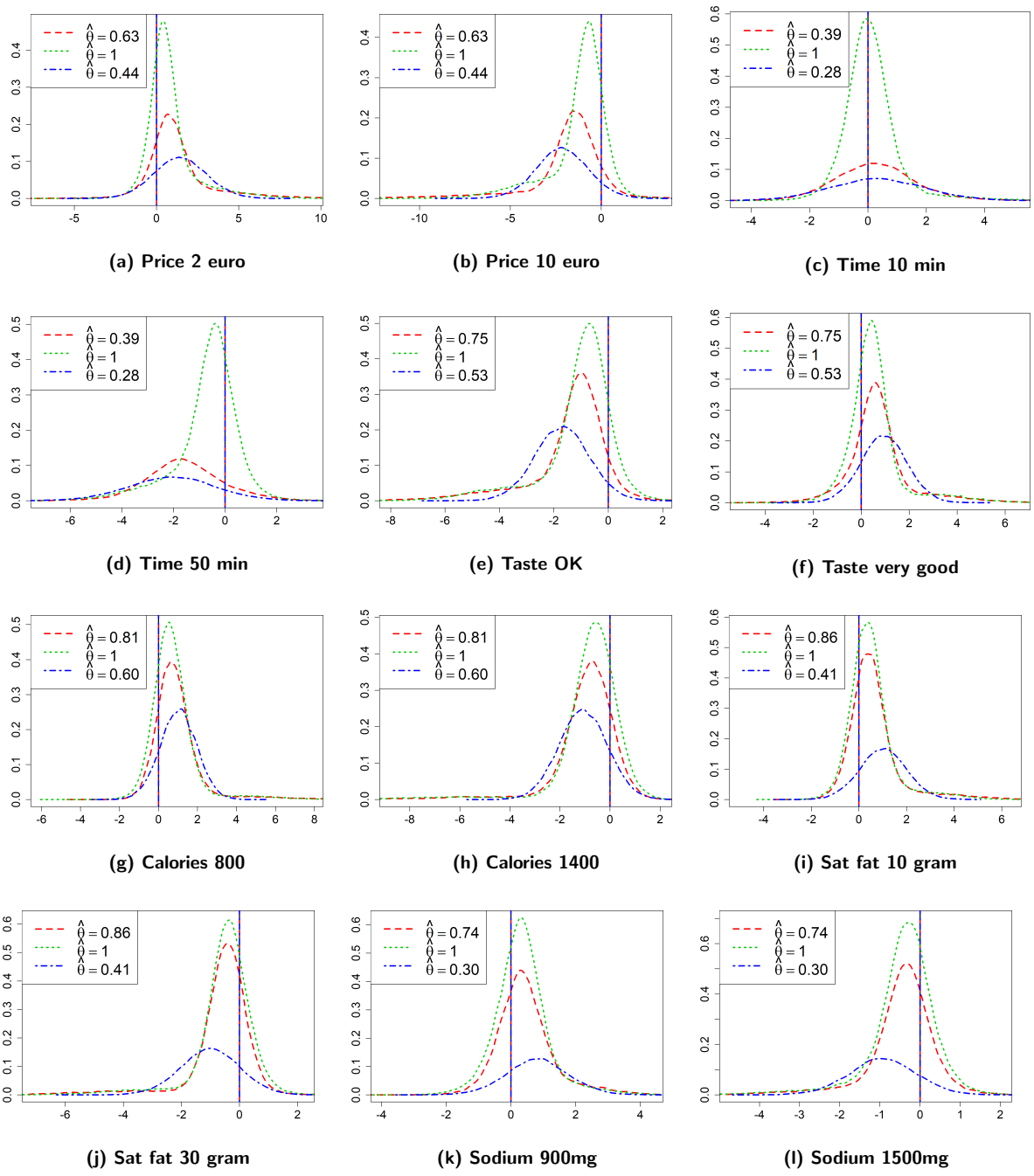
Baseline levels are price 6 euro, cooking time 30 minutes, taste good, calories 1100, saturated fat 20 gram, and sodium 1200 mg.

HVS-DPM

DPM

HVS-M

Figure 6: Posterior marginal density of β_{ik} for respondent group 3.



(a) Price 2 euro **(b) Price 10 euro** **(c) Time 10 min**

(d) Time 50 min **(e) Taste OK** **(f) Taste very good**

(g) Calories 800 **(h) Calories 1400** **(i) Sat fat 10 gram**

(j) Sat fat 30 gram **(k) Sodium 900mg** **(l) Sodium 1500mg**

Baseline levels are price 6 euro, cooking time 30 minutes, taste good, calories 1100, saturated fat 20 gram, and sodium 1200 mg.

HVS-DPM
DPM
HVS-M

5.2. Out-of-sample performance

To further evaluate the performance of our approach for modeling heterogeneous responses, we look at the forecasting performance for the empirical dataset on food choice. We use predictive Bayes factors to compare the performance using our approach as compared to using the standard DP mixture (DPM) and the single multivariate normal with heterogeneous variable selection (HVS-M). For this purpose, we rerun the sampler to obtain posterior samples based on a subset of the observations: for each individual we randomly remove two observations. Based on these observations left out, we compute log predictive Bayes factors of our approach versus one of the two competing approaches.[12] A log predictive Bayes factor greater than log(3) indicates that there is sufficient evidence to favor our approach (Kass & Raftery, 1995). A log Bayes factor smaller than log(1/3) indicates that there is sufficient evidence to favor the alternative approach.

For posterior results, we again use 60,000 simulations after 40,000 burn-in draws and keep every 10^{th} draw. We repeat this exercise ten times per respondent group, using different randomly chosen forecasting samples to increase the robustness of the results to the selection of the forecasting sample. For each respondent group, we report the averages across these ten replications, and the percentages of replications for which the log Bayes factor is positive.

The results for the log predictive Bayes factors are given in Table 7. The averages are all positive, indicating that our approach leads to a better forecasting performance than the alternative approaches. Our approach clearly stands out as compared to the HVS-M approach with a single multivariate normal that was proposed by Gilbride et al. (2006). The log Bayes factors are much larger than zero for the majority of forecasting samples, and the averages exceed log(3) (≈ 1.10) for all three respondent groups. Hence, for this dataset on

[12]The log predictive Bayes factor of our approach against one of the alternative approaches is computed by subtracting the predictive log-likelihood of the alternative approach from the predictive log-likelihood from our approach. The predictive log-likelihood is approximated using the posterior samples:

$$\sum_{i=1}^{N} \log \left[\frac{1}{S} \sum_{s=1}^{S} \prod_{t \in \mathcal{T}_i^*} \Pr[Y_{it} = y_{it} | \beta_i^{(s)}] \right], \tag{14}$$

where S is the number of draws of the MCMC sampler after burn-in, $\mathcal{T}_i^*$ is the set of observations for unit i that was left out of the training sample and $\beta_i^{(s)}$ is the s^{th} posterior draw of β_i which can be computed directly using the s^{th} posterior draws of δ_i and τ_i.

Table 7: Log predictive Bayes factors for our approach (HVS-DPM) against two alternative approaches (DPM and HVS-M). Averaged across ten different forecasting samples. Also reports the percentage of samples for which the log Bayes factor is positive.

	HVS-DPM against DPM		HVS-DPM against HVS-M	
DGP	Mean	$\% > 0$	Mean	$\% > 0$
Group 1	1.13	70%	12.26	90%
Group 2	0.85	60%	8.16	70%
Group 3	6.39	80%	10.80	100%

food choices, there is sufficient evidence of non-normality in the distribution of preferences. These findings indicate that allowing for flexible cross-sectional heterogeneity via a mixture of multivariate normals is important for understanding and predicting choice behavior.

Our approach also compares favorably to the DPM approach without heterogeneous variable selection, although the results are less overwhelming for respondent groups 1 and 2. A possible reason for the relative small difference in predictive performance could be that quite some individuals have strong preferences (> 2) for one attribute or another, as indicated by the heavy tails. In the multinomial logit model, such attributes will dominate the choice predictions, making it less important to accurately estimate the preferences close to zero. The final respondent group, for which the predictive performance of our approach clearly stands out, was the group which had to make decisions on the largest number of attributes (six) without obtaining objective information on the health attributes. For this group, allowing for heterogeneous variable selection substantially improves the predictive performance.

6. Conclusion

In this paper, we develop a general method for heterogeneous variable selection in Bayesian nonlinear panel data models. We allow for flexible cross-sectional heterogeneity by letting the model's unit-specific parameters follow a Dirichlet process mixture of multivariate normals. Our main contribution is that we augment the DP mixture with heterogeneous variable selection. This allows modeling the possibility that subsets of units are unaffected by certain

variables, as may be present in applications as diverse as health treatments, choice situations, macroeconomics, and operations research. We develop our approach for nonlinear panel data models including multinomial logit and probit models, count models, exponential models, among many others. Finally, we develop an efficient Bayesian MCMC sampler to allow for inference for datasets with up to 50 or 100 explanatory variables.

We illustrate the model with a Monte Carlo study and an empirical application. For illustration, we consider a multinomial logit model as this model is the focus of most literature on heterogeneous variable selection. In the Monte Carlo study we find that our approach is able to capture both complex forms of continuous cross-sectional heterogeneity — such as skewness and multimodality — as well as heterogeneous variable selection. A 'standard' DP mixture cannot capture heterogeneous variable selection. Instead of a spike at zero, this approach generally allocates probability mass to a relatively large region around zero, depending on the shape of the continuous heterogeneity. In the empirical application, we consider responses to a discrete choice experiment on food choices. We find substantial evidence of attribute non-attendance and non-normality of the continuous heterogeneity. In particular, the continuous heterogeneity seems skewed. These findings indicate the usefulness of our approach in practice.

A limitation of the proposed approach is the use of a conjugate prior for the components' means and covariance matrices. Although this prior is advantageous for estimation, it may be unrealistic as the prior on the component's mean directly depends on the component's covariance matrix. This implies that the marginal prior on the mean is tighter when the corresponding variance is small. If the conjugacy of the prior would be relaxed, it is required to draw the component membership indicators with a Metropolis-Hastings step instead of a Gibbs step. This could dramatically increase computation time due to worse mixing properties of the resulting MCMC sampler.

We note three interesting venues for future research. First, one can allow for correlated variable selection. This could be incorporated, for example, by allowing for a different membership probability per combination of variables selected and putting a (Dirichlet) prior on these membership probabilities. Second, the model can be generalized to allow for time-varying parameters, including time-varying variable selection. In choice situations, this could

model changing preferences of individuals, or learning and fatigue effects. Finally, the non-linear (univariate) panel data model can be extended to multivariate outcomes. This would require inference on the correlations across outcomes.

References

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian non-parametric problems. *The Annals of Statistics*, 1152–1174.

Cai, B., & Dunson, D. (2005). Variable selection in nonparametric random effects models. *Technical report, Department of Statistical Science, Duke University.*

Campbell, D., Hensher, D. A., & Scarpa, R. (2011). Non-attendance to attributes in environmental choice analysis: A latent class specification. *Journal of Environmental Planning and Management*, *54*(8), 1061–1076.

Collins, A. T., Rose, J. M., & Hensher, D. A. (2013). Specification issues in a generalised random parameters attribute nonattendance model. *Transportation Research Part B: Methodological*, *56*, 234–253.

Cook, S. R., Gelman, A., & Rubin, D. B. (2006). Validation of software for Bayesian models using posterior quantiles. *Journal of Computational and Graphical Statistics*, *15*(3), 675–692.

Dunson, D. B., Herring, A. H., & Engel, S. M. (2008). Bayesian selection and clustering of polymorphisms in functionally related genes. *Journal of the American Statistical Association*, *103*(482), 534–546.

Fan, W., & Bouguila, N. (2013). Variational learning of a Dirichlet process of generalized Dirichlet distributions for simultaneous clustering and feature selection. *Pattern Recognition*, *46*(10), 2754–2769.

Ferguson, T. S. et al. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2*(4), 615–629.

Frühwirth-Schnatter, S., & Frühwirth, R. (2010). Data augmentation and MCMC for binary and multinomial logit models. *Statistical Modelling and Regression Structures* (pp. 111–132). Springer.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*(423), 881–889.

George, E. I., & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, 339–373.

Geweke, J. (2004). Getting it right: Joint distribution tests of posterior simulators. *Journal of the American Statistical Association*, *99*(467), 799–804.

Gilbride, T. J., Allenby, G. M., & Brazell, J. D. (2006). Models for heterogeneous variable selection. *Journal of Marketing Research*, *43*(3), 420–430.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hensher, D. A., Collins, A. T., & Greene, W. H. (2013). Accounting for attribute non-attendance and common-metric aggregation in a probabilistic decision process mixed multinomial logit model: A warning on potential confounding. *Transportation*, *40*(5), 1003–1020.

Hensher, D. A., & Greene, W. H. (2010). Non-attendance and dual processing of common-metric attributes in choice analysis: A latent class specification. *Empirical Economics*, *39*(2), 413–426.

Hess, S., Stathopoulos, A., Campbell, D., O'Neill, V., & Caussade, S. (2013). It's not that I don't care, I just don't care very much: Confounding between attribute non-attendance and taste heterogeneity. *Transportation*, *40*(3), 583–607.

Hole, A. R. (2011). A discrete choice model with endogenous attribute attendance. *Economics Letters*, *110*(3), 203–205.

Hole, A. R., Kolstad, J. R., & Gyrd-Hansen, D. (2013). Inferred vs. stated attribute non-attendance in choice experiments: A study of doctors' prescription behaviour. *Journal of Economic Behavior & Organization*, *96*, 21–31.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kim, S., Dahl, D. B., & Vannucci, M. (2009). Spiked Dirichlet process prior for Bayesian multiple hypothesis testing in random effects models. *Bayesian Analysis (Online)*, *4*(4), 707.

Kim, S., Tadesse, M. G., & Vannucci, M. (2006). Variable selection in clustering via Dirichlet process mixture models. *Biometrika*, *93*(4), 877–893.

Kim, S., DeSarbo, W. S., & Fong, D. K. (2018). A hierarchical Bayesian approach for examining heterogeneity in choice decisions. *Journal of Mathematical Psychology*, *82*, 56–72.

Koç, H., & van Kippersluis, H. (2017). Thought for food: Nutritional information and educational disparities in diet. *Journal of Human Capital*, *11*(4), 508–552.

Korobilis, D. (2013). Bayesian forecasting with highly correlated predictors. *Economics Letters*, *118*(1), 148–150.

MacLehose, R. F., Dunson, D. B., Herring, A. H., & Hoppin, J. A. (2007). Bayesian methods for highly correlated exposure data. *Epidemiology*, *18*(2), 199–207.

Manski, C. F. (1977). The structure of random utility models. *Theory and Decision*, *8*(3), 229.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105–142). Academic Press: New York.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Mitchell, T. J., & Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, *83*(404), 1023–1032.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *9*(2), 249–265.

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

Roberts, G. O., Gelman, A., Gilks, W. R., et al. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, *7*(1), 110–120.

Roberts, G. O., Rosenthal, J. S. et al. (2001). Optimal scaling for various Metropolis-Hastings algorithms. *Statistical Science*, *16*(4), 351–367.

Rossi, P. (2015). Bayesm: Bayesian inference for marketing/micro-econometrics, 2012. *URL http://CRAN. R-project. org/package= bayesm. R package version*, 2–2.

Rossi, P. (2014). *Bayesian non-and semi-parametric methods and applications*. Princeton University Press.

Scarpa, R., Gilbride, T. J., Campbell, D., & Hensher, D. A. (2009). Modelling attribute non-attendance in choice experiments for rural landscape valuation. *European Review of Agricultural Economics*, *36*(2), 151–174.

Tang, X., Xue, F., & Qu, A. (2020). Individualized multi-directional variable selection. *Journal of the American Statistical Association*, (forthcoming).

Wang, C., & Blei, D. M. (2009). Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process. *Advances in Neural Information Processing Systems*, 1982–1989.

Wu, F., Han, Y., Liu, X., Shao, J., Zhuang, Y., & Zhang, Z. (2012). The heterogeneous feature selection with structural sparsity for multimedia annotation and hashing: A survey. *International Journal of Multimedia Information Retrieval*, *1*(1), 3–15.

Yang, M. (2012). Bayesian variable selection for logistic mixed model with nonparametric random effects. *Computational Statistics & Data Analysis*, *56*(9), 2663–2674.

Yu, G., Huang, R., & Wang, Z. (2010). Document clustering via Dirichlet process mixture model with feature selection. *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 763–772.

Zhao, L., Hu, Q., & Wang, W. (2015). Heterogeneous feature selection with multi-modal deep neural networks and sparse group lasso. *IEEE Transactions on Multimedia*, *17*(11), 1936–1948.

A. MCMC sampler

In this section, we develop the MCMC sampler for our nonlinear panel data model with heterogeneous variable selection in Equations (1)-(10). In summary, the model is given by

$$\begin{aligned} Y_{it}|\beta_i, \gamma &\sim f(g(x_{it}, \beta_i, z_{it}, \gamma)), \\ \beta_{ik}|\tau_{ik}, \lambda_{ik} &= \tau_{ik}\lambda_{ik}, \end{aligned}$$

with variable selection priors

$$\begin{aligned} \tau_{ik} &\in \{\kappa, 1\}, \\ \Pr[\tau_{ik} = 1|\theta_k] &= \theta_k, \\ \theta_k &\sim Beta(a, b), \end{aligned}$$

DP mixture priors

$$\begin{aligned} \lambda_i|\{\pi_q\}_q, \{\mu_q\}_q, \{\Sigma_q\}_q &\sim \sum_{q=1}^{\infty} \pi_q MVN(\mu_q, \Sigma_q) \\ \pi_q &= \eta_q \prod_{r=1}^{q-1}(1-\eta_r), \quad \eta_q \sim Beta(1, \alpha), \\ \mu_q|\Sigma_q &\sim MVN(\mu_0, d^{-1}\Sigma_q), \\ \Sigma_q &\sim IW(\nu, \nu\upsilon I), \end{aligned}$$

and finally

$$\gamma \sim MVN(\gamma_0, \Sigma_\gamma).$$

The hyperparameters α, μ_0, d, ν, υ, γ_0, Σ_γ, a and b, are assumed fixed. The sampler can be easily extended to allow for priors on these hyperparameters.

The MCMC sampler is given by

(1) $c_i|c_{-i}, \lambda_i, \{\mu_q\}_q, \{\Sigma_q\}_q$ for $i = 1, ..., N$, (Gibbs, multinomial),

(2) $\mu_q, \Sigma_q|\{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N$ for every unique q in $\{c_1, ..., c_N\}$:

($2a$) $\Sigma_q|\{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N$ (Gibbs, inverse Wishart),

($2b$) $\mu_q|\{\lambda_i\}_{i=1}^N, \{c_i\}_{i=1}^N, \Sigma_q$ (Gibbs, multivariate normal),

(3) $\lambda_{ik}, \tau_{ik}|y_i, \lambda_{i,-k}, \tau_{i,-k}, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta_k, \gamma$ for $i = 1, ..., N$, and $k = 1, ..., K_x$ *(in random order)*:

($3a$) $\lambda_{ik}|y_i, \lambda_{i,-k}, \tau_{i,-k}, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta_k, \gamma$ (M-H, random walk),

($3b$) $\tau_{ik}|y_i, \lambda_i, \tau_{i,-k}, \theta_k, \gamma$ (Gibbs, Bernoulli),

(4) $\theta_k|\{\tau_{ik}\}_{i=1}^N$ for $k = 1, ..., K_x$, (Gibbs, Beta),

(5) $\gamma|\{y_i\}_{i=1}^N, \{c_i\}_{i=1}^N, \{\lambda_i\}_{i=1}^N, \{\tau_i\}_{i=1}^N$ (M-H, random walk),

In this sampler, we jointly draw μ_q and Σ_q, and we jointly draw λ_{ik} and τ_{ik}.

Two remarks on this sampler. First, in case some of the variables should be simultaneously selected, and thus $K_x^* < K_x$ (see Section 3, right before Section 3.1), step 3 should be slightly altered to loop over all $k = 1, ..., K_x^*$ and, per k, to jointly draw $\{\lambda_{il}, \tau_{il}\}$ over all l for which $D_{l,k}^* = 1$. Second, in case K_x is really small, say $K_x < 5$, the MCMC sampler could be more efficient when λ_i and τ_i are jointly drawn over all variables instead of per variable k. In this case, step 3 can be replaced by step 3* below

(3*) $\lambda_i, \tau_i\ |y_i, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta, \gamma$ for $i = 1, ..., N$:

($3a^*$) $\lambda_i|y_i, c_i, \mu_{c_i}, \Sigma_{c_i}, \theta, \gamma$ (M-H, random walk),

($3b^*$) $\tau_i|y_i, \lambda_i, \theta, \gamma$ (Gibbs, Multinomial).

The starting values are generated as follows. First, we set the number of components to 10, and generate the component membership indicators c_i from a multinomial distribution with 10 outcomes, each with equal probability. Second, the components means μ_q are set to a vector of zeroes, and the component covariance matrices Σ_q to an identity matrix. Third, to draw λ_i and γ, we first compute the maximum likelihood estimates of the parameters of

the corresponding model with homogeneous responses and no variable selection. Then, for each individual, we take the λ_i that optimizes a weighted log-likelihood function.[13] Fourth, we set $\theta_k = 0.95$ for $k = 1, ..., K_x$. Finally, we draw τ_{ik} by first drawing a r_{ik} from a Bernoulli distribution with parameter θ_k and then setting τ_{ik} to 1 when $r_{ik} = 1$ and to κ otherwise.

A.1. Draw c_i

We use algorithm 2 from Neal (2000) to sample c_i from the full conditional posterior. Let n_c denote the number of units in component c, and $n_{c,-i}$ denote the number of units in component c if we would not count unit i. Let $\mathcal{Q}_i$ be the current set of distinct components if we would not count unit i. That is, $\mathcal{Q}_i$ consists of the distinct components in $\{c_1, ..., c_N\} \backslash \{c_i\}$. Let Q_i be the size of the set $\mathcal{Q}_i$. We draw c_i from a multinomial distribution with $Q_i + 1$ outcomes. The first Q_i possible outcomes are the objects in $\mathcal{Q}_i$, the final component is a new component. The corresponding probabilities are given by

$$\Pr[c_i = q | c_{-i}, \lambda_i, \{\mu_q\}_q, \{\Sigma_q\}_q] = \begin{cases} \frac{n_{q,-i} f(\lambda_i|\mu_q, \Sigma_q)}{\sum_{r \in \mathcal{Q}_i} n_{r,-i} f(\lambda_i|\mu_r, \Sigma_r) + \alpha \int f(\lambda_i|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma}, & \text{if } q \in \mathcal{Q}_i, \\ \frac{\alpha \int f(\lambda_i|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma}{\sum_{r \in \mathcal{Q}_i} n_{r,-i} f(\lambda_i|\mu_r, \Sigma_r) + \alpha \int f(\lambda_i|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma}, & \text{if } q \notin \mathcal{Q}_i, \end{cases}$$

where $f(\lambda_i|\mu_q, \Sigma_q)$ is the density of a multivariate normal distribution with mean μ_q and covariance matrix Σ_q evaluated at λ_i, $f(\mu, \Sigma)$ is the prior density of a μ and Σ based on the prior distribution in Equation (7)-(8) and the marginal density of λ_i is given by

$$f(\lambda_i) = \int f(\lambda_i|\mu, \Sigma) f(\mu, \Sigma) d\mu d\Sigma = \left(\frac{d}{\pi(d+1)}\right)^{K_x/2} \frac{\Gamma_{K_x}((\nu+1)/2)}{\Gamma_{K_x}(\nu/2)} \frac{|\nu \upsilon I|^{\nu/2}}{|\hat{S}_i|^{(\nu+1)/2}},$$

[13] The weighted log-likelihood function that is optimized over λ_i is similar to the one used in Rossi (2015) for the MNL and is given by

$$0.9 * \log f(y_i|\lambda_i, \gamma) + 0.1 * \frac{T_i}{\sum_i T_i}(-0.5 * z'z), \tag{15}$$

where $f(y_i|\lambda_i, \gamma)$ is the likelihood function of observing y_i conditional on $\beta_i = \lambda_i$ and γ, and $z = L(\lambda_i - \hat{\lambda})$ where $\hat{\lambda}$ is the pooled maximum likelihood estimate of λ, and L is the Cholesky decomposition of the negative Hessian of the pooled log-likelihood function at the maximum likelihood estimates.

where Γ_K is the multivariate Gamma function, $|\cdot|$ denotes the determinant and $\hat{S}_i$ is the scale matrix of the distribution of Σ conditional on λ_i as given by

$$\hat{S}_i = \nu\upsilon I + (\lambda_i - \hat{\mu}_i)(\lambda_i - \hat{\mu}_i)' + d(\mu_0 - \hat{\mu}_i)(\mu_0 - \hat{\mu}_i)', \tag{16}$$

where $\hat{\mu}_i$ is the mean of the distribution of μ conditional on λ_i as given by

$$\hat{\mu}_i = \frac{d\mu_0 + \lambda_i}{d+1}. \tag{17}$$

For these derivations, we use the conjugacy of the normal-inverse Wishart prior on μ and Σ.

When in the multinomial distribution we draw a new component $c_i \notin \mathcal{Q}_i$, we also need to draw a new component mean μ_{c_i} and covariance matrix Σ_{c_i}. These are drawn from their posterior. For this purpose, we first draw Σ_{c_i} conditional on λ_i, and then μ_{c_i} conditional on Σ_{c_i} and λ_i. That is, we draw Σ_{c_i} from an inverse Wishart distribution with $\nu + 1$ degrees of freedom and scale matrix $\hat{S}_i$. Next, we draw μ_{c_i} from a multivariate normal distribution with mean $\hat{\mu}_i$ and covariance matrix $(d+1)^{-1}\Sigma_{c_i}$.

A.2. Draw Σ_q and μ_q

We can jointly draw Σ_q and μ_q conditional on $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$ by first drawing Σ_q conditional on $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$ and then drawing μ_q conditional on Σ_q, $\{\lambda_i\}_{i=1}^N$ and $\{c_i\}_{i=1}^N$, for $q = 1, ...Q$.

We draw Σ_q from an inverse Wishart distribution with degrees of freedom $\nu + N_q$ and scale matrix

$$\hat{S}_q = \nu\upsilon I + \sum_{i=1}^{N} I[c_i = q](\lambda_i - \hat{\mu}_q)(\lambda_i - \hat{\mu}_q)' + d(\mu_0 - \hat{\mu}_q)(\mu_0 - \hat{\mu}_q)' \tag{18}$$

where N_q is the number of units in component q, and

$$\hat{\mu}_q = \frac{d\mu_0 + \sum_{i=1}^{N} I[c_i = q]\lambda_i}{d + N_q}. \tag{19}$$

Next, we draw μ_q from a multivariate normal distribution with mean $\hat{\mu}_q$ and covariance

matrix $(d + N_q)^{-1}\Sigma_q$.

A.3. Draw λ_{ik}

We use a random walk Metropolis-Hastings step to draw λ_{ik} conditional on y_i, $\lambda_{i,-k}$, $\tau_{i,-k}$, c_i, μ_{c_i}, Σ_{c_i}, θ_k, and γ. Conditional on $\lambda_{i,-k}$ and $\tau_{i,-k}$, we know $\beta_{i,-k}$. Moreover, given $\lambda_{i,-k}$, μ_{c_i} and Σ_{c_i}, the prior for λ_{ik} is a univariate normal distribution with mean $\tilde{\mu}_{\lambda_{ik}}$ and variance $\tilde{\sigma}^2_{\lambda_{ik}}$ given by

$$\tilde{\mu}_{\lambda_{ik}} \equiv E[\lambda_{ik}|\lambda_{i,-k}, \mu_{c_i}, \Sigma_{c_i}] = \mu_{c_i,k} + \Sigma_{c_i,k,-k}\Sigma^{-1}_{c_i,-k,-k}(\lambda_{i,-k} - \mu_{c_i,-k}), \tag{20}$$

$$\tilde{\sigma}^2_{\lambda_{ik}} \equiv \text{Var}\left(\lambda_{ik}|\lambda_{i,-k}, \Sigma_{c_i}\right) \;= \Sigma_{c_i,kk} - \Sigma_{c_i,k,-k}\Sigma^{-1}_{c_i,-k,-k}\Sigma_{c_i,-kk}, \tag{21}$$

where $\Sigma_{c,k,-k}$ refers to the k^{th} row of Σ and all columns except for the k^{th}.

The candidate for λ_{ik} is drawn from the normal distribution

$$\lambda^*_{ik} \sim N(\lambda_{ik}, \rho^2_{\lambda,ik}\tilde{\sigma}^2_{\lambda_{ik}}), \tag{22}$$

where $\rho_{\lambda,ik}$ is a parameter to be tuned such the acceptance rate is about 0.44 (Roberts et al., 1997, Roberts, Rosenthal, et al., 2001). Tuning is performed during the burn-in MCMC iterations. The candidate is accepted with probability

$$\min\left[1, \frac{f(y_i|\lambda^*_{ik}, \beta_{i,-k}, \theta_k, \gamma)f(\lambda^*_{ik}|\mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k})}{f(y_i|\lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma)f(\lambda_{ik}|\mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k})}\right], \tag{23}$$

where the likelihood contribution conditional on λ_{ik} and $\beta_{i,-k}$ is given by

$$f(y_i|\lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma) = \sum_{\tilde{\tau}_{ik}\in\{\kappa,1\}} f(y_i, \tilde{\tau}_{ik}|\lambda_{ik}, \beta_{i,-k}, \theta_k, \gamma), \tag{24}$$

$$= \sum_{\tilde{\tau}_{ik}\in\{\kappa,1\}} \Pr[\tau_{ik} = \tilde{\tau}_{ik}|\theta_k] f(y_i|\lambda_{ik}, \tilde{\tau}_{ik}, \beta_{i,-k}, \gamma), \tag{25}$$

$$= \sum_{\tilde{\tau}_{ik}\in\{\kappa,1\}} \theta_k^{I[\tilde{\tau}_{ik}=1]}(1-\theta_k)^{I[\tilde{\tau}_{ik}=\kappa]} f(y_i|\tilde{\beta}_i, \gamma), \tag{26}$$

$$= \sum_{\tilde{\tau}_{ik}\in\{\kappa,1\}} \theta_k^{I[\tilde{\tau}_{ik}=1]}(1-\theta_k)^{I[\tilde{\tau}_{ik}=\kappa]} \left(\prod_{t=1}^{T_i} f(y_{it}|\tilde{\beta}_i, \gamma)\right), \tag{27}$$

where $\tilde{\beta}_i$ has k^{th} element $\tilde{\tau}_{ik}\lambda_{ik}$ and $f(y_{it}|\tilde{\beta}_i, \gamma)$ is the likelihood contribution of observation t of unit i conditional on $\tilde{\beta}_i$ and γ given in Equation (1). For the prior density of λ_{ik} we have that

$$f(\lambda_{ik}|\mu_{c_i}, \Sigma_{c_i}, \lambda_{i,-k}) \propto \exp\left\{-\frac{1}{2}\frac{(\lambda_{ik} - \tilde{\mu}_{\lambda_{ik}})^2}{\tilde{\sigma}^2_{\lambda_{ik}}}\right\}. \tag{28}$$

In case λ_i should be drawn jointly over all variables k, the candidate should be a multivariate normal distribution and the tuning parameter ρ_λ should be tuned such to obtain an acceptance rate of about 0.234 (Roberts et al., 1997, Roberts, Rosenthal, et al., 2001).

A.4. Draw τ_{ik}

We draw τ_{ik} conditional on y_i, λ_i, $\tau_{i,-k}$, θ_k and γ using a Bernoulli distribution. The conditional probability that τ_{ik} is equal to 1 is given by

$$\Pr[\tau_{ik} = 1|y_i, \lambda_i, \tau_{i,-k}, \theta_k, \gamma] = \frac{\Pr[\tau_{ik} = 1|\theta_k] f(y_i|\lambda_{ik}, \tau_{ik} = 1, \beta_{i,-k}, \gamma)}{\sum_{\tilde{\tau}_{ik} \in \{\kappa, 1\}} \Pr[\tau_{ik} = \tilde{\tau}_{ik}|\theta_k] f(y_i|\lambda_{ik}, \tilde{\tau}_{ik}, \beta_{i,-k}, \gamma)} \tag{29}$$

$$= \frac{\theta_k \left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \lambda_{ik}, \beta_{i,-k}, \gamma)\right)}{(1-\theta_k)\left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \kappa\lambda_{ik}, \beta_{i,-k}, \gamma)\right) + \theta_k \left(\prod_{t=1}^{T_i} f(y_{it}|\beta_{ik} = \lambda_{ik}, \beta_{i,-k}, \gamma)\right)} \tag{30}$$

where the likelihood contribution of observation t of unit i conditional on β_i and γ is given in Equation (1). Hence, we can draw a r_{ik} from a Bernoulli distribution with the probability in Equation (30). Then, we obtain a draw of τ_{ik} by setting τ_{ik} equal to 1 when $r_{ik} = 1$ and equal to κ when $r_{ik} = 0$.

A.5. Draw θ_k

We can directly draw θ_k conditional on $\{\tau_{ik}\}_{i=1}^N$ from the Beta distribution

$$\theta_k|\{\tau_{ik}\}_{i=1}^N \sim Beta\left(a + \sum_i I[\tau_{ik} = 1], b + \sum_i I[\tau_{ik} = \kappa]\right), \tag{31}$$

for $k = 1, ..., K_x$.

A.6. Draw γ

We use a random walk Metropolis-Hastings step to draw γ conditional on $\{y_i\}_{i=1}^N$, $\{c_i\}_{i=1}^N$, $\{\lambda_i\}_{i=1}^N$, and $\{\tau_i\}_{i=1}^N$. First notice that given $\{\lambda_i\}_{i=1}^N$, and $\{\tau_i\}_{i=1}^N$, we know $\{\beta_i\}_{i=1}^N$. At the s^{th} draw, the candidate for γ, γ^*, is drawn from

$$\gamma^* \sim MVN(\gamma^{(s-1)}, \rho_\gamma^2 \Sigma_{c_i}), \tag{32}$$

where $\gamma^{(s-1)}$ is the current draw for γ, and $\rho_{\lambda,i}$ is a parameter to be tuned such the acceptance rate is about 0.234 (Roberts et al., 1997, Roberts, Rosenthal, et al., 2001). The acceptance probability is given by

$$\min\left[1, \frac{f(y_i|\gamma^*, \beta_i) f(\gamma^*)}{f(y_i|\gamma^{(s-1)}, \beta_i) f(\gamma^{(s-1)})}\right], \tag{33}$$

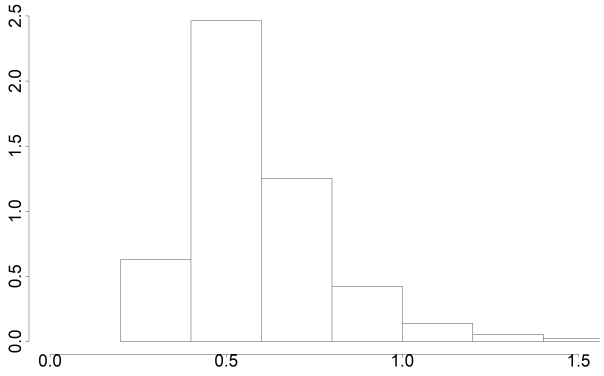
where

$$f(y_i|\gamma, \beta_i) = \prod_{t=1}^{T_i} f(y_{it}|\beta_i, \gamma), \tag{34}$$

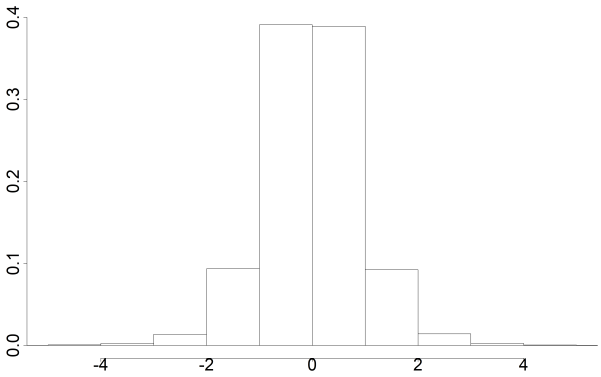
and $f(\gamma)$ is the prior density of γ. In case γ^* is not accepted, we set $\gamma^{(s)} = \gamma^{(s-1)}$.

B. Histograms of priors

Figure 7: Priors μ_q and Σ_q

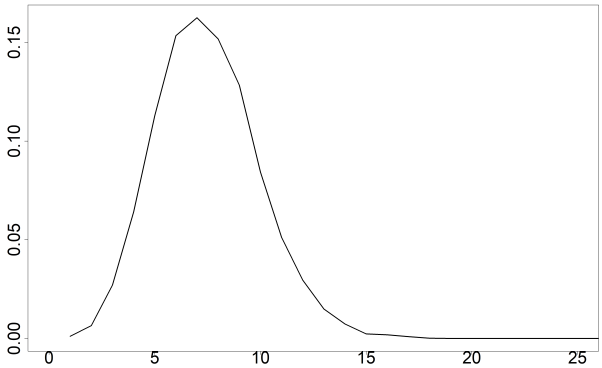


(a) Prior standard deviation $\sqrt{Diag(\Sigma_q)}$ in Monte Carlo study and empirical applications. This is the marginal density of the square root of a variance from the diagonal of a covariance matrix based on an $IW(\nu, \nu\upsilon I)$ distribution with $K = 3$, $\nu = K + 5$, and $\upsilon = 0.2$.



(b) Prior mean for μ_q, marginalized over Σ_q, in Monte Carlo study and empirical applications. This is the marginal density based on a $MVN(0, 0.5^{-1}\Sigma_q)$ prior for μ, $K = 3$, and the prior $\Sigma_q \sim IW(\nu, \nu\upsilon I)$ with $\nu = K + 5$, and $\upsilon = 0.2$.

Figure 8: Implied prior on number of components ($N = 1,000$ and $\alpha = 1$)



C. Hit rates Monte Carlo study

Table 8: Percentage of units for which the posterior draw of β_{ik} falls within $-\epsilon \leq \beta_{ik} \leq \epsilon$ for multiple values of ϵ (averaged over Monte Carlo replications, draws and variables).

	(1) All				(2) True $-\epsilon \leq \beta_{ik} \leq \epsilon$			(3) True $\beta_{ik} < -\epsilon$ or $\beta_{ik} > \epsilon$			(4) True $\beta_{ik} = 0$		
ϵ	True	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M	HVS-DPM	DPM	HVS-M
DGP 1													
0.00	10	7	0	10	24	0	36	6	0	8	24	0	36
0.10	13	11	5	15	29	13	38	8	4	11	31	12	42
0.20	17	14	10	19	36	25	44	10	7	14	39	25	48
0.30	20	19	16	23	43	36	49	13	11	17	46	36	54
0.40	23	24	22	28	49	45	55	16	15	20	54	47	60
0.50	28	29	29	33	55	53	59	19	19	24	61	57	66
0.75	43	46	47	48	67	68	68	30	32	32	77	76	78
1.00	64	64	65	62	78	79	76	40	41	38	88	88	87
1.50	88	86	86	85	92	92	90	42	44	52	97	97	97
2.00	95	94	94	96	96	96	97	59	60	81	99	99	99
2.50	98	98	98	99	98	98	100	78	77	95	100	100	100
DGP 2													
0.00	0	2	0	2	-	-	-	2	0	2	-	-	-
0.10	3	5	3	6	16	13	17	4	3	6	-	-	-
0.20	7	8	6	11	26	24	28	6	5	9	-	-	-
0.30	10	11	10	15	35	33	37	8	7	13	-	-	-
0.40	14	15	15	20	42	41	44	11	10	16	-	-	-
0.50	19	21	20	26	47	46	50	14	14	20	-	-	-
0.75	37	39	39	41	61	61	61	26	26	29	-	-	-
1.00	59	59	59	57	74	75	72	36	36	35	-	-	-
1.50	86	83	84	83	90	91	88	39	39	51	-	-	-
2.00	94	93	93	96	96	96	97	58	59	81	-	-	-
2.50	98	98	98	99	98	98	99	79	79	95	-	-	-
DGP 3													
0.00	25	16	0	19	33	0	40	11	0	13	33	0	40
0.10	28	21	8	24	39	15	45	13	5	15	40	15	46
0.20	30	25	16	28	46	29	51	16	10	18	47	29	53
0.30	33	30	24	33	52	42	57	18	15	21	54	42	59
0.40	36	35	32	38	59	53	62	22	21	24	61	53	65
0.50	40	41	40	43	64	61	67	25	26	27	67	63	70
0.75	53	56	58	56	75	76	75	36	38	35	81	81	82
1.00	70	71	72	69	83	84	81	43	45	40	90	91	90
1.50	90	88	89	88	93	93	92	45	47	53	98	98	98
2.00	95	95	95	97	97	97	98	62	63	81	100	100	100
2.50	98	98	98	100	99	99	100	79	79	94	100	100	100
DGP 4													
0.00	10	14	0	14	25	0	26	12	0	12	25	0	26
0.10	17	20	9	20	34	16	35	17	8	17	34	15	35
0.20	24	26	19	26	45	32	45	20	15	20	43	29	44
0.30	31	33	28	33	54	46	54	24	20	24	51	43	52
0.40	39	40	38	40	61	57	61	27	26	27	60	54	60
0.50	47	48	47	48	68	66	68	30	31	30	67	64	67
0.75	66	66	68	66	80	81	80	40	43	40	82	82	82
1.00	81	81	82	81	88	89	88	51	54	51	91	92	91
1.50	97	96	96	96	97	97	97	77	76	77	99	99	99
2.00	100	100	99	100	100	99	100	94	93	94	100	100	100
2.50	100	100	100	100	100	100	100	99	99	100	100	100	100

Results for four different groups:
(1) all units,
(2) units for which true β_{ik} falls within interval,
(3) units for which true β_{ik} does not fall within interval,
(4) units for which true $\beta_{ik} = 0$ ($\tau_{ik} = 0$).