How to generalize from a hierarchical model?

Max J. Pachali, Peter Kurz & Thomas Otter

Quantitative Marketing and Economics
QME

ISSN 1570-7156
Volume 18
Number 4

Quant Mark Econ (2020) 18:343-380
DOI 10.1007/s11129-020-09226-7

Springer

Quantitative Marketing and Economics (2020) 18: 343–380
https://doi.org/10.1007/s11129-020-09226-7

How to generalize from a hierarchical model?

Max J. Pachali[1] · **Peter Kurz**[2] · **Thomas Otter**[3]

Received: 17 April 2018 / Accepted: 1 April 2020 / Published online: 17 May 2020

Abstract

Models of consumer heterogeneity play a pivotal role in marketing and economics, specifically in random coefficient or mixed logit models for aggregate or individual data and in hierarchical Bayesian models of heterogeneity. In applications, the inferential target often pertains to a population beyond the sample of consumers providing the data. For example, optimal prices inferred from the model are expected to be optimal in the population and not just optimal in the observed, finite sample. The population model, random coefficients distribution, or heterogeneity distribution is the natural and correct basis for generalizations from the observed sample to the market. However, in many if not most applications standard heterogeneity models such as the multivariate normal, or its finite mixture generalization lack economic rationality because they support regions of the parameter space that contradict basic economic arguments. For example, such population distributions support positive price coefficients or preferences against fuel-efficiency in cars. Likely as a consequence, it is common practice in applied research to rely on the collection of individual level mean estimates of consumers as a representation of population preferences that often substantially reduce the support for parameters in violation of economic expectations. To overcome the choice between relying on a mis-specified heterogeneity distribution and the collection of individual level means that fail to measure heterogeneity consistently, we develop an approach that facilitates the formulation of more economically faithful heterogeneity distributions based on prior constraints. In the common situation where the heterogeneity distribution comprises both constrained and unconstrained coefficients (e.g., brand and price coefficients), the choice of subjective prior parameters is an unresolved challenge. As a solution to this problem, we propose a marginal-conditional decomposition that avoids the conflict between wanting to be more informative about constrained parameters and only weakly informative about unconstrained parameters. We show how to efficiently sample from the implied posterior and illustrate the merits of our prior as well as the drawbacks of relying on

R-code for running the MCMC sampler developed in this paper is available at https://github.com/mpachali/How-to-Generalize-Hierarchical-Model-Replication.

✉ Max J. Pachali
m.j.pachali@tilburguniversity.edu

Extended author information available on the last page of the article.

means of individual level preferences for decision-making in two illustrative case studies.

Keywords Discrete choice · Bayesian inference · Market simulation · Constrained hierarchical prior

JEL Classification C01 · C11 · C35 · C33 · M31

1 Introduction

Models of consumer heterogeneity play a pivotal role in marketing and economics. Typical applications are random coefficients or mixed logit models for aggregate or panel data (e.g., Revelt and Train 1998 and Train 2009), and hierarchical Bayesian models. Influential applications of these models involve inference from household scanner panel data or from discrete choice experiments (e.g., Allenby and Lenk 1994, Rossi et al. 1996, Allenby et al. 1998, Dubé et al. 2010, and Sawtooth 2013). In most applications, the inferential target pertains to a population beyond the sample of consumers providing the data for model calibration. For example, pricing, product design, or product line decisions informed by the sample data through the model are expected to be optimal in the population and not just in the observed, finite sample. The population model, the heterogeneity or random coefficients distribution is the natural and correct basis for generalizations from the observed sample of consumers or respondents to the market. The fact that inferences about parameters of this distribution are consistent in the sample size (N), even if the number of observations contributed by each consumer (T) is very small, makes this approach attractive from a statistical perspective.

Unfortunately, standard population distributions often lack economic rationality. For example, Reiss and Wolak (2007) remark that the estimated distribution of marginal utility of fuel economy in Berry et al. (1995) suggests that about half of consumers in the car market dislike fuel economy. As another example, Dubé et al. (2008, 2010) find support for positive price coefficients in the inferred heterogeneity distribution. Such economically unreasonable characterizations of consumer heterogeneity prevent meaningful counterfactual predictions from the model. As an obvious example, models that support positive price coefficients in the inferred heterogeneity distribution preclude model based price optimization.

While a completely theory driven specification of heterogeneity distributions appears to be beyond reach, some authors argue in favor of theory driven constraints in the population distribution (e.g., Boatwright et al. 1999 and Allenby et al. 2014). The goal is a heterogeneity model that is maximally flexible regarding some aspects of the population distribution, but deterministically constrained by economic theory regarding other aspects of this distribution. This paper builds on this idea and develops it further.

In applications, a prior understanding of preferences in the population often suggest a large number of sign and order restrictions, for example: that the price parameter in an indirect utility function is negative or that consumers prefer a more

fuel efficient to a less fuel efficient car, everything else equal. So called constrained parameter problems are relevant across academic fields and a body of literature dealt with this topic. Gelfand et al. (1992) provide an overview of how to impose sign and order constraints based on truncated distributions using Gibbs sampling. Allenby et al. (1995) introduce this approach into marketing in the context of individual level conjoint analysis. Boatwright et al. (1999) develop a sampler in the spirit of Gelfand et al. (1992), but for a hierarchical sales response regression model.

However, sign and order restrictions in models of heterogeneity still present unresolved challenges. In principle, one could adopt truncated normal distributions that implement prior constraints as outlined in Gelfand et al. (1992) for heterogeneity distributions. However, as we show below, any truncated distribution of heterogeneity leads to a so called "doubly intractable" inference problem. The log-normal prior avoids this difficulty. The basic idea of using log-normal distributions to implement sign and order constraints is not new. For example, Allenby et al. (2014) use the exponential transformation, $\beta_p = -\exp(\beta_p^*)$ with $\beta_p^* \in \mathbb{R}$ distributed according to a hierarchical normal mixture prior, to enforce that the model has zero support for positive price coefficients. In this specification, the problem is that β_p^* is measured on the log scale and standard diffuse subjective prior settings imply absurdly large and small values of transformed coefficients β_p (e.g., Allenby et al. 2014).[1] In the common situation where the heterogeneity distribution thus comprises both constrained and unconstrained coefficients, the choice of subjective prior parameters is an unresolved challenge.

As a solution to this problem we propose a marginal-conditional decomposition that avoids the conflict between wanting to be more subjectively informative about constrained parameters and only weakly informative about unconstrained parameters. We show that this decomposition is important whenever the heterogeneity distribution comprises a mix of constrained and unconstrained coefficients, e.g., brand and price coefficients. Our decomposition applies both to the fully parametric multivariate normal setting as well as to its semi-parametric generalizations. In addition, we show how to efficiently sample from the implied posterior building on the likelihood based pre-tuning of proposal densities in Rossi et al. (2005).

Finally, we contrast profit implications of relying on the inferred population distribution to an ad-hoc approach that approximates heterogeneity using means of individual level coefficients. This latter approach is still common in applied academic and industry research. It is ad-hoc because if fails to measure heterogeneity consistently, distorting inference towards the population mean. As a consequence, markets will misleadingly appear too homogeneous, translating into too little product differentiation and too much price competition in counterfactual calculations. A side-effect of this distortion is a reduction of sign and order violations in the approximated heterogeneity distribution that likely contributed to the popularity of this ad-hoc approach.

[1] See also Peter Rossi's vignette on the impact of prior specifications in constrained parameter problems: https://cran.r-project.org/web/packages/bayesm/vignettes/Constrained_MNL_Vignette.html. Accessed: 20th November 2019.

In a nutshell the goal of this paper is to facilitate the formulation of more economically faithful hierarchical prior distributions of heterogeneity for better market simulators and improved counterfactual calculations. We thereby hope to broaden the applicability of models of heterogeneity, and to convince applied academic and industry researchers to abandon market simulators built on means of individual level preferences. The remainder of the paper proceeds as follows: Section 2 formally introduces different ways of generalizations from hierarchical Bayesian models and discusses implications for market simulation. In Section 3 we develop the hierarchical prior formulation and in Section 4 we discuss efficient sampling of individual level coefficients. Section 5 then investigates the relative performance of the proposed approach using simulated data. Sections 6 and 7 report the results from two empirical illustrations based on household scanner panel data on purchases of fresh hen's eggs (Kotschedoff and Pachali 2020) and data from a discrete-choice experiment on tablet PCs. Finally, we summarize and discuss results in Section 8.

2 Different ways of generalizations and market simulations

Different ways of generalizing from hierarchical models to consumer preferences, choices, and market shares in the target population are best illustrated in a decision theoretic framework. For this purpose, and without loss of generality, we abstract away from competition and fixed costs, and assume constant marginal prices and costs in the following. If the decision-maker knew the distribution of preferences in the population denoted as $p(\beta|\tau)$, he would choose the action $a \in A$ that maximizes profits $\int \pi(a,\beta)\ p(\beta|\tau)\ d\beta = \mathbb{E}_{\beta|\tau}[\pi(a,\beta)] = \pi(a)$ by solving the following maximization problem:

$$\max_{a \in A} \left\{ \pi(a) \propto (P(a) - C(a)) \int \mathrm{MS}(a,\beta) p(\beta|\tau)\, d\beta \right\} \tag{1}$$

Here $\mathrm{MS}(a,\beta)$ is the market share from action a and preference β, as implied by a choice model, $C(a)$ denotes marginal costs associated with action a, and $P(a)$ the marginal price, which may itself constitute an action; thus $(P(a) - C(a))$ is the contribution margin. Finally, the proportionality results from ignoring the market size.

Because the preference distribution in the population is generally unknown, the decision-maker forms an expectation about profits based on data $Y = \left(y_1 \ \ldots \ y_i \ \ldots \ y_N \right)$, where y_i is the T_i-vector of observations from individual i in the sample, and based on prior assumptions about the choice model underlying $\mathrm{MS}(a,\beta)$, the distribution of preferences in the population $p(\beta|\tau)$, and the parameters τ in this distribution. He then maximizes the posterior expected profit:

$$\hat{\pi}(a) = \mathbb{E}_{\beta|Y}[\pi(a,\beta)] \propto (P(a) - C(a)) \int \mathrm{MS}(a,\beta) p(\beta|\tau) p(\tau|Y)\, d(\beta,\tau) \tag{2}$$

This estimator of expected profits entirely relies on posterior knowledge of the hierarchical prior distribution. We thus refer to this approach as "generalizing based on the hierarchical prior". It is easily computed to an arbitrary degree of precision

based on MCMC draws from the posterior distribution $p(\tau|Y)$ coupled with draws from the hierarchical prior distribution $p(\beta|\tau)$. However, because it entirely relies on the posterior of the hierarchical prior, all prior parametric assumptions will come to bear. If, for example, the hierarchical prior supports positive and negative price coefficients as in a normal distribution, the posterior of the hierarchical prior will necessarily—and may substantially—support positive price coefficients. The problem may persist even if the data reliably locate *all* individual specific posterior price coefficient distributions in the negative domain. The reason is that the best normal approximation matches the first and second moment of the distribution to be fitted, which may result in substantial support for positive coefficients even if *all* coefficients to be fitted are negative.

To mitigate the extrapolation of parametric assumptions in directions that violate economic theory, market simulators often rely on the collection of individual level posterior mean estimates $\{\hat{\beta}_i\}_{i=1}^N$ where $\hat{\beta}_i = \int \beta_i \, p(\beta_i|Y, y_i) d\beta_i$ —the shrinkage of individual level posterior means to the population mean in general reduces the number of sign and order violations, albeit at the expense of severely inconsistent inferences about heterogeneity. Expected profits from action a are then estimated as:

$$\hat{\pi}(a) \propto (P(a) - C(a)) \frac{1}{N} \sum_{i=1}^{N} \mathrm{MS}(a, \hat{\beta}_i) \tag{3}$$

However, as we illustrated in Appendix A.1, this estimator that aggregates optimal, in the sense of a bias-variance trade-off, individual level estimates, itself fails optimality criteria and is inconsistent no matter how large the sample of consumers N, as long as individual level likelihoods are not perfectly informative about individual level preferences. In practice, individual level likelihoods tend to be diffuse, which motivates hierarchical models in the first place.

A third estimator of expected profits from action a builds on the collection of individual level posterior distributions. We refer to this form of generalization as lower level model non smoothed (n.s.) because it relies on the lower, individual level models, but does not summarize individual level posteriors to estimates.

$$\hat{\pi}(a) \propto (P(a) - C(a)) \frac{1}{N} \sum_{i=1}^{N} \int \mathrm{MS}(a, \beta_h) p(\beta_h|y_i, \tau) p(\tau|Y) \, d(\beta_h, \tau) \tag{4}$$

The difference between this estimator and that defined in Eq. 2 is that y_i is used both to inform the posterior $p(\tau|Y)$ and the prediction to new consumers' preferences in $p(\beta_h|y_i, \tau)$. When individual level posterior distributions essentially degenerate to a point because of highly informative individual level likelihoods, the estimator in Eq. 4 converges to that defined in Eq. 3. When individual level posterior distributions come from diffuse individual level likelihoods, as usual, the estimator in Eq. 4 will be very similar to that in Eq. 2. Thus, parametric assumptions in the hierarchical prior distributions will be similarly influential. Consistent with these assessments, we only find negligible differences between generalizations based on the posterior of the hierarchical prior and lower level model n.s. in the empirical applications discussed below.

What way of generalization should we use for market simulation in practice? Every trained Bayesian analyst will point out the inconsistency associated with relying on the collection of individual level posterior means. Such an analyst knows that posterior predictive preference distributions as defined in Eqs. 2 and 4 allow for consistent inference (in N), however conditional on functional form assumptions.

However, because standard parametric and semi-parametric assumptions such as multivariate normal or its finite mixture generalization violate basic economic intuition in many applications, consistency conditional on these assumptions is not too helpful. Thus, many applied researchers and practitioners opt for generalizations, i.e., market simulation based on the collection of individual level posterior means (Eq. 3) that often substantially reduce the share of sign and order violations. We aim to overcome the choice between relying on the posterior of a mis-specified hierarchical prior and the collection of individual level posterior means that fail to measure heterogeneity, by showing how to specify more economically faithful hierarchical prior distributions based on prior constraints. The goal is a hierarchical prior that both is maximally flexible regarding some aspects of the population distribution of preferences, and deterministically constrained by theory regarding other aspects of this distribution.

3 Sign and order constraints

Sign and order constraints dogmatically express prior knowledge about the support of a distribution, e.g., that the price parameter in an indirect utility function is negative or that a consumer prefers a more fuel efficient to a less fuel efficient car for sure, everything else equal. So called constrained parameter problems are relevant across academic fields and a body of literature dealt with this topic. Gelfand et al. (1992) provide an overview of how to impose sign and order constraints based on truncated distributions using Gibbs sampling. Allenby et al. (1995) introduce this approach into marketing in the context of individual level conjoint analysis. Boatwright et al. (1999) develop a sampler in the spirit of Gelfand et al. (1992), but for a hierarchical sales response regression model.

However, the implementation of sign and order restrictions in hierarchical Bayesian models is still without a generally accepted solution. In principle, one could adjust the sampler outlined by Gelfand et al. (1992) to hierarchical settings. However, as we show next, any truncation applied to the prior (and hence to the posterior) of individual level coefficients in a hierarchical setting leads to a so called "doubly intractable" inference problem in the hierarchical prior. Doubly intractable problems are characterized by a normalization constant that depends on target parameters (e.g., Möller et al. 2006 and Murray et al. 2006). Consider the following truncated normal hierarchical prior for consumers' demand parameters:

$$p(\beta|\bar{\beta}, V_\beta) = \frac{\varphi(\beta|\bar{\beta}, V_\beta)}{\mathbb{Z}(\bar{\beta}, V_\beta)}\mathbf{1}(\beta \in \mathbb{R}_c^k), \tag{5}$$

where $\mathbb{R}_c^k$ denotes the truncation region of a k-dimensional demand parameter vector β, φ denotes the multivariate normal density and $\mathbb{Z}(\bar{\beta}, V_\beta)$ the corresponding normalizing constant:

$$\mathbb{Z}(\bar{\beta}, V_\beta) = \int_{\mathbb{R}_c^k} \varphi(\beta|\bar{\beta}, V_\beta) d\beta \tag{6}$$

The conditional posterior distribution of parameters indexing the hierarchical prior then becomes:

$$p(\bar{\beta}, V_\beta|\{\beta_i\}) \propto \prod_{i=1}^{N} \frac{\varphi(\beta_i|\bar{\beta}, V_\beta)}{\mathbb{Z}(\bar{\beta}, V_\beta)} \mathbf{1}(\beta_i \in \mathbb{R}_c^k) p(\bar{\beta}, V_\beta), \tag{7}$$

where $p(\bar{\beta}, V_\beta)$ denotes the subjective prior for hierarchical prior parameters. Equation 7 is an example of a doubly intractable inference problem because even after dropping the normalization constant $\int \left(\prod_{i=1}^{N} \frac{\varphi(\beta_i|\bar{\beta}, V_\beta)}{\mathbb{Z}(\bar{\beta}, V_\beta)} \mathbf{1}(\beta_i \in \mathbb{R}_c^k) p(\bar{\beta}, V_\beta) \right) d\{\bar{\beta}, V_\beta\}$ of the posterior giving rise to the proportionality, we are left with the intractable expression $\mathbb{Z}(\bar{\beta}, V_\beta)$. This expression normalizes the multivariate normal density to the region of support defined by $\mathbb{R}_c^k$ and cannot be dropped because it depends on target parameters $\bar{\beta}$ and V_β.[2]

As a consequence of truncation, we loose the convenience of conditionally conjugate updates of hierarchical prior parameters $\bar{\beta}$ and V_β regardless of what subjective prior distributions we employ. More generally, all estimation and sampling techniques that require the evaluation of the conditional "likelihood" $p(\{\beta_i\}|\bar{\beta}, V_\beta) = \prod_{i=1}^{N} \frac{\varphi(\beta_i|\bar{\beta}, V_\beta)}{\mathbb{Z}(\bar{\beta}, V_\beta)}$, including standard Metropolis-Hastings sampling, are hamstrung by the intractability of $\mathbb{Z}(\bar{\beta}, V_\beta)$.[3] Boatwright et al. (1999) propose to numerically approximate $\mathbb{Z}(\bar{\beta}, V_\beta)$ at each MCMC iteration using the GHK algorithm (Hajivassiliou et al. 1996). While this seems reasonable in their application that involves sign constraints on at most four parameters in a model with five parameters in total, numerical approximations will be problematic in the high-dimensional parameter spaces, potentially involving a multiplicity of constraints that have become common in applications more recently.

The log-normal hierarchical prior avoids this difficulty. The basic idea of using log-normal distributions to implement sign and order constraints is not new. For example, Allenby et al. (2014) use the exponential transformation, $\beta_p = -\exp(\beta_p^*)$

[2]Note that without truncation, i.e., when $\mathbb{R}_c^k = \mathbb{R}^k$, $\mathbb{Z}(\bar{\beta}, V_\beta) = 1$ for all regular $\bar{\beta}$ and V_β.

[3]Some researchers simply ignore $\mathbb{Z}(\bar{\beta}, V_\beta)$ in the update of upper level parameters, i.e., use standard updates based on $p(\bar{\beta}, V_\beta|\{\beta_i\}) \propto \prod_{i=1}^{N} \varphi(\beta_i|\bar{\beta}, V_\beta) p(\bar{\beta}, V_\beta)$. This "solution" results in an incoherent model in the sense that data generating parameters may not be recovered, even from infinitely large samples.

with $\beta_p^* \in \mathbb{R}$ and distributed according to a hierarchical normal mixture prior, to enforce that the model has zero support for positive price coefficients. In this specification, the problem is that β_p^* is measured on the log scale and standard diffuse subjective prior settings imply absurdly large and small values of transformed coefficients β_p (e.g., Allenby et al. 2014).

Thus, the problem is how to specify differentially informative subjective priors for constrained coefficients and unconstrained coefficients. The standard Normal-Inverse-Wishart (NIW) subjective prior for means and covariance matrices in the hierarchical prior distribution is limited in this regard—mostly because the prior concentration of the IW-prior is controlled by a single parameter (the prior degrees of freedom also known as the prior shape).

Next, we present a solution to this problem that re-parameterizes the hierarchical prior. Our contributions in this context are, first, a marginal-conditional decomposition of the hierarchical prior distribution that enables the analyst to be differentially informative about the distribution of constrained and unconstrained parameters in the population a priori[4], and second, the generalization of the pre-tuning of proposal densities in Rossi et al. (2005) to this hierarchical prior.

The proposed marginal-conditional decomposition becomes essential whenever the hierarchical prior comprises both constrained and unconstrained parameters such as e.g., in simple hierarchical choice models that feature brand coefficients and a price coefficient. The proposed generalization of pre-tuned proposal densities (Rossi et al. 2005) is particularly important in high dimensional models that feature a multiplicity of constraints.

3.1 Marginal-conditional decomposition

Our hierarchical prior starts with a standard normal distribution.[5] Unconstrained coefficients have a normal hierarchical prior while sign and order constraints are imposed through exponential transformations of normal variates resulting in log-normally distributed coefficients. Vice versa, we can log-transform from sign and order constrained parameters that enter the likelihood to unconstrained, a priori conditionally normally distributed variates. We formulate subjective priors over this unconstrained space but use a marginal-conditional decomposition to implement vastly different subjective priors for parameters that are exponentiated and those that are not.

We denote $g : \mathbb{R}^k \to \mathbb{R}_c^k$ as the function that maps normally distributed variates β_i^* to sign and order constrained coefficients β_i that enter multinomial likelihoods explaining individual choice data y_i. We distinguish k_c "constrained" coefficients

[4] See McCulloch et al. (2000) for another example of specifying flexible priors for covariance matrices.

[5] We focus on the single component normal model to minimize notational clutter. The generalizations to mixtures of normals is straightforward.

β_i^{*c}, i.e., coefficients to be transformed to obey sign and order constraints, and k_{uc} unconstrained coefficients β_i^{*uc} in the hierarchical prior.

$$\begin{aligned} y_i|g(\beta_i^*) &\sim MNL\left(y_i|g(\beta_i^*)\right) \\ \beta_i^* &\sim N\left(\bar{\beta}^*, V_{\beta^*}\right), \text{ or} \\ \begin{pmatrix} \beta_i^{*c} \\ \beta_i^{*uc} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_c^* \\ \mu_{uc}^* \end{pmatrix}, \begin{pmatrix} V_{\beta_{11}^*} & V_{\beta_{12}^*} \\ V_{\beta_{21}^*} & V_{\beta_{22}^*} \end{pmatrix}\right) \end{aligned} \tag{8}$$

With the goal of formulating rather different subjective priors for the parameters governing the distribution of β_i^{*c} and β_i^{*uc}, we re-express the multivariate normal distribution in Eq. 8 in the form of a multivariate regression model that regresses unconstrained coefficients β_i^{*uc} on "constrained" coefficients β_i^{*c}:

$$B^{*uc} = \begin{pmatrix} \iota & B^{*c} \end{pmatrix} \begin{pmatrix} z' \\ \Gamma \end{pmatrix} + U \qquad vec(U') \sim N(0, I_N \otimes \Sigma) \tag{9}$$

Here, B^{*uc} and B^{*c} are matrices with k_{uc} and k_c columns, respectively, and N rows each, collecting unconstrained and "constrained" coefficients from individuals in the sample, and ι is a $(N \times 1)$-vector of 1's; Γ is a $(k_c \times k_{uc})$ matrix of regression coefficients, z a column vector of intercept coefficients of length k_{uc}, and Σ is the $(k_{uc} \times k_{uc})$ conditional variance-covariance of unconstrained coefficients in the population.

The first two moments of the distribution of "constrained" coefficients are obtained from yet another multivariate regression model that regresses "constrained" coefficients on a vector of constants:

$$B^{*c} = \iota(\mu_c^*)' + U_{V^*} \qquad vec(U'_{V^*}) \sim N(0, I_N \otimes V^*) \tag{10}$$

Here, ι is again a $(N \times 1)$-vector of 1's and V^* is the marginal variance-covariance matrix of constrained coefficients. The multivariate regression models in Eqs. 9 and 10 imply the following re-parameterization of the joint distribution of β_i^* from Eq.8:

$$\beta_i^* \sim N\left(\begin{pmatrix} \mu_c^* \\ \Gamma'\mu_c^* + z \end{pmatrix}, \begin{pmatrix} V^* & V^*\Gamma \\ \Gamma'(V^*)' & \Gamma'V^*\Gamma + \Sigma \end{pmatrix}\right) \tag{11}$$

The advantage of the re-parameterization in Eq. 11 relative to the more standard parameterization in Eq. 8 is that we can now specify arbitrarily informative subjective priors for the hierarchical prior distribution of "constrained" coefficients, i.e., for the parameters μ_c^* and V^* without restricting the prior of unconstrained coefficients. That is, if we a priori set V^* to a "small" covariance matrix, we can nevertheless elect to be minimally informative about the distribution of unconstrained parameters through Σ. Coupled with weakly informative priors for Γ and z, neither the correlation between "constrained" and unconstrained nor the marginal mean of unconstrained coefficients is directly affected by informative prior specifications for μ_c^* and V^*.

However, the role of the prior on Γ in the implied prior for the covariance of unconstrained coefficients (see the lower right block of the covariance matrix in Eq. 11) requires additional discussion. A priori, an increasing number of constrained coefficients coupled with a diffuse prior on Γ implies a marginal prior for the variance of

Table 1 Quantiles of marginal prior densities for a constrained coefficient with informative and standard weakly informative subjective priors

	Informative	Weakly informative
1%	−1.934E+03	−2.576E+10
25%	−8.977E+00	−1.054E+03
50%	−9.914E-01	−1.049E+00
75%	−1.132E-01	−1.031E-03
99%	−5.098E-04	−3.951E-11

unconstrained coefficients that may appear as favoring larger variances. In this context, it is important to keep in mind that the variance contribution through Γ is through the covariance between "constrained" and unconstrained coefficients (see the upper right and lower left block of the covariance matrix in Eq. 11). Thus, the prior implication of large marginal variances of unconstrained coefficients stems from "mixing" over strong and qualitatively different (positive or negative) dependencies between constrained and unconstrained coefficients. However, strong dependence between "constrained" and unconstrained coefficients constitutes an extremely informative hierarchical prior. Hence, "mixing" over strong and qualitatively different (positive or negative) dependencies between constrained and unconstrained coefficients is not a possibility a posteriori, not even in small data sets. For example, even smallish data sets will enforce a choice between the two highly informative opposites of strong positive and strong negative dependence between a constrained and an unconstrained coefficient. In sum, large variances of unconstrained coefficients through Γ a posteriori result from strong dependence between "constrained" and unconstrained coefficients as per the likelihood.

Before going into more detail about suggested subjective choices, we illustrate the problem of formulating sensible priors for constrained coefficients in the smallest possible example where $\beta_i = -\exp(\beta_i^*)$, $\beta_i^* \sim N(\bar{\beta}^*, V_{\beta^*})$. Here, the subjective prior is on parameters $\bar{\beta}^*$ and V_{β^*} in the normal distribution that generates β_i^*. Under what is widely considered a weakly informative subjective prior setting[6] for $\bar{\beta}^*$ and V_{β^*}, we obtain that a priori 25% of the constrained coefficients $\{\beta_i\}$ are larger than $-.001$, i.e., very close to zero, and another 25% are smaller than -1054 (see the right column in Table 1).

This concentration of mass in the tails of the prior is undesirable and counter to what one would expect from a weakly informative prior for β_i. The prior for β_i in the column on the left in Table 1 has lower (upper) quartiles of -8.977 ($-.113$) and appears to be much more reasonable for, say, the population distribution of price coefficients in a heterogeneous multinomial logit model. However, this marginal prior distribution requires subjective priors for $\bar{\beta}^*$ and V_{β^*} discussed next that in most applications would be considered unduly informative as a prior for unconstrained coefficients where $\beta_i = \beta_i^*$.

[6] We follow Rossi et al. (2005) in the specification of a weakly informative subjective prior setting. Specifically, for a single parameter problem and based on the paramterization in Eq. 12: $A_{\mu_c^*} = 0.01$, $\nu_{V^*} = 6$ as well as $\bar{V}^* = \nu_{V^*}$.

We use the fully conjugate prior for (Γ_z, Σ), where $\Gamma_z := \left(z, \Gamma'\right)'$, and the conditionally conjugate prior for (μ_c^*, V^*):

$$\begin{aligned} p\left(\Gamma_z, \Sigma\right) &= p\left(\Gamma_z|\Sigma\right) p(\Sigma) \\ \gamma_z|\Sigma &\sim N(\bar{\gamma}_z, \Sigma \otimes A_{\Gamma_z}^{-1}),\ \gamma_z := vec\left(\Gamma_z\right) \\ \Sigma &\sim IW(\nu_\Sigma, \bar{\Sigma}) \text{ and} \\ p(\mu_c^*, V^*) &= p(\mu_c^*)p(V^*), \\ \mu_c^* &\sim N(\bar{\mu}_c^*, A_{\mu_c^*}^{-1}) \\ V^* &\sim IW(\nu_{V^*}, \bar{V}^*) \end{aligned} \tag{12}$$

The conditionally conjugate prior for (μ_c^*, V^*) enables the researcher to directly express prior beliefs about the distribution of "constrained" coefficients in the population. We set $\bar{\mu}_c^* = \left(0 \ \ldots \ 0\right)'$, $A_{\mu_c^*} = 0.1 I_{k_c}$, $\nu_{V^*} = k_c + 15$ as well as $\bar{V}^* = 0.5\nu_{V^*} I_{k_c}$, where I_{k_c} is the identity matrix of dimension $k_c \times k_c$ (cf. Allenby et al. 2014). Especially the choice of prior degrees of freedom ν_{V^*}, i.e., the shape parameter in the IW prior for V^*, would be considered unduly informative as a default value in the context of only unconstrained parameters. However, our marginal-conditional decomposition of the hierarchical prior enables the analyst to be arbitrarily informative about the hierarchical prior for "constrained" coefficients, essentially without affecting the marginal hierarchical prior for unconstrained coefficients.

The fully conjugate prior for (Γ_z, Σ) adjusts the influence of the subjective prior on Γ_z as a function of the conditional variance-covariance Σ, which is desirable in situations without much prior knowledge. We use standard weakly informative, "barely proper" priors for parameters in the conditional hierarchical prior of unconstrained coefficients, $\bar{\gamma}_z,\ A_{\Gamma_z},\ \nu_\Sigma,\ \bar{\Sigma}$.

Our marginal-conditional decomposition corresponds to the directed acyclic graph in Fig. 1 which shows that the hierarchical prior for "constrained coefficients", (μ_c^*, V^*), and that of unconstrained coefficients, (Γ_z, Σ), are independent conditional on draws of "constrained" coefficients, B^{*c}. This conditional independence relationship gives rise to a Gibbs-sampler for the two-stage update of parameters in the hierarchical prior:

1. $\beta_i^*|(\mu_c^*, V^*), (\Gamma_z, \Sigma), y_i,\ i = 1, \ldots, N$

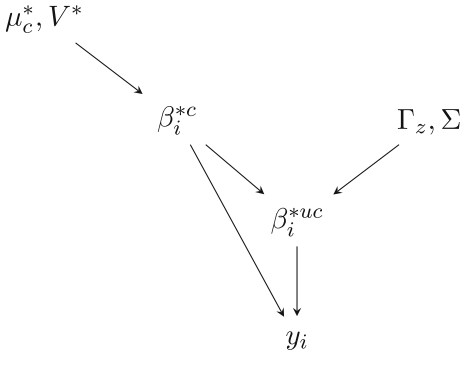

Fig. 1 Marginal-conditional decomposition DAG

2. $\{\Gamma_z, \Sigma\} | B^{*uc}, B^{*c}$
3. $\left\{\mu_c^*, V^*\right\} | B^{*c}$

In step 1, we use a random walk Metropolis-Hastings (RW-MH) step to draw individual level parameters $\left\{\beta_i^*\right\}$ based on multinomial logit likelihoods, similar to Rossi et al. (2005). However, as described in detail in the following Section 4, we need to account for the change of variables in $g : \mathbb{R}^k \to \mathbb{R}_c^k$ when tuning the MH-proposal using information from the likelihood. In step 2, we use a Gibbs-sampler to update Γ_z and Σ, i.e., parameters in a fully conjugate multivariate regression model, conditional on both "constrained" and unconstrained coefficients and subjective prior parameters (omitted for simplification). Step 3 employs another Gibbs-step to update (μ_c^*, V^*), i.e., parameters in a conditionally conjugate multivariate regression model, conditional on "constrained" coefficients and subjective prior parameters. Appendix A.2 details the posterior distributions associated with steps two and three.

4 Efficient MH-sampling

Next we discuss efficient sampling of individual level part worth coefficients $\{\beta_i^*\}$ based on pre-tuned proposal densities in a MH-sampler conditional on draws of hierarchical prior parameters (Rossi et al. 2005). Our algorithmic implementation is for a MNL model at the individual level, but the approach obviously generalizes to other likelihoods. The pre-tuning in Rossi et al. (2005) employs a normal approximation to the likelihood. The MNL-likelihood information about $\{\beta_i\}$ can be computed in closed form. However, our hierarchical prior is on the distribution of $\{\beta_i^*\}$; therefore, we need to account for the change-of-variables in $g : \mathbb{R}^k \to \mathbb{R}_c^k$.

Following Rossi et al. (2005), we specify the proposal density of the RW-MH sampler as follows

$$\beta_i^{*cand} \sim N\left(\beta_i^{*r}, c^2\left(H_i^* + (V_{\beta^*}^r)^{-1}\right)^{-1}\right), \tag{13}$$

where $r \in \{1, \ldots, R\}$ is the r-th iteration of the MCMC chain, c denotes a fixed scaling factor and H_i^* is the Hessian information about β_i^* in individual i's data, evaluated at the maximum of the following fractional likelihood:

$$l_i^{\text{fract}}\left(\{y_i\}_{i=1}^N | g(\beta_i^*)\right) = MNL\left(y_i | g(\beta_i^*)\right)^{1-w} \; MNL\left(\{y_i\}_{i=1}^N | g(\beta_i^*)\right)^{w(T_i/\bar{T})} \tag{14}$$

This fractional likelihood is defined as a w-weighted combination of the individual specific likelihood and the likelihood of a model that pools all observations, where T_i is the number of choice observations from individual i and $\bar{T}$ is the total number of choices made by all individuals in the calibration sample.

At the maximizing value $\check{\beta}_i$ we can straightforwardly transform to $\check{\beta}_i^*$ by standard maximum likelihood theory. We obtain the corresponding H_i^* in Eq. 15, taking advantage of the closed form expression for the information about β_i, denoted H_i,

from individual i's choices in the MNL model, and accounting for the change of variable to a first order approximation.[7]

$$H_i^* \approx \left(J_g\right)' H_i J_g \tag{15}$$

Here J_g is the $k \times k$ Jacobian of the function $g(\beta_i^*)$ that maps conditional normally distributed variates β_i^* to their sign and order constrained counterparts β_i. H_i and J_g are evaluated at $\check{\beta}_i$ and $g^{-1}(\check{\beta}_i) = \check{\beta}_i{}^*$ respectively, i.e., at the parameter value that maximizes the fractional likelihood in Eq. 14.

Appendix A.4 illustrates the value of the proposed tuning in the MH-upate of β_i^* in a small simulation that only involves choices of one individual. We find that the proposed tuning results in a sampler that is on average about 3.7 times more efficient than that using a simpler and more standard tuning (see Table 15). We note that these differences can magnify substantially in a hierarchical setting.

5 Simulation study

Next we illustrate the benefits of our proposed marginal-conditional decomposition in the presence of sign and order constraints using simulations. First, we compare prior distributions in the prototypical setting that combines constrained and unconstrained coefficients. Second, we analyze the posterior from simulated data under different priors and elaborate on the numerical properties of the proposed methodology.

5.1 Drawing from prior distributions

Suppose a hypothetical setting with two attributes $A1$ and $A2$ at two levels $L1$ and $L2$ each, yielding four possible product configurations. Both levels of the first attribute provide positive utility to every consumer, and its second level is weakly preferred to the first, again by all consumers. To reflect these sign and order restrictions, we denote the respective coefficients as $\{\beta_{+,i}\}$ and $\{\beta_{++,i}\}$, where $i = 1, \ldots, N$ indexes simulated consumers. Preferences for the levels of the second attribute are heterogeneous but without a uniform prior direction or ordering, such as e.g., the preferences for colors or flavors in applications. We denote the respective coefficients as $\{\beta_{uc_1,i}\}$ and $\{\beta_{uc_2,i}\}$. The price coefficient is negative. We thus have the following set of constraints for every consumer $i = 1, \ldots, N$:

$$\begin{aligned} \beta_{+,i}, \beta_{++,i} &\geq 0 \\ \beta_{++,i} &\geq \beta_{+,i} \\ \beta_{p,i} &\leq 0 \end{aligned} \tag{16}$$

First, we compare (implied) marginal priors for coefficients $\beta = g(\beta^*)$ based on the marginal-conditional decomposition in Eq. 11 and a more standard parameteri-

[7] Appendix A.3 provides the derivation of the exact Hessian of transformed coefficients. We found improvements from using the exact Hessian to be small in applications, relative to the first order approximation in Eq. 15.

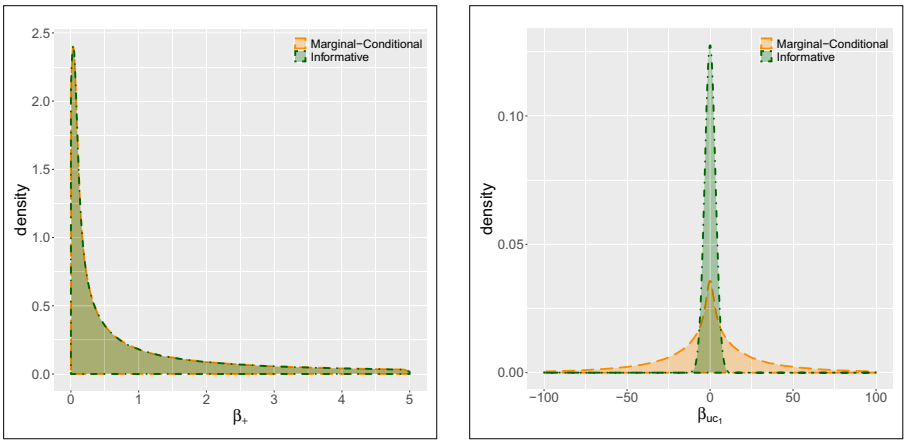

Fig. 2 Marginal prior distributions of β_+ (left panel) and β_{uc_1} (right panel) using the marginal-conditional decomposition and the standard formulation

zation (Eq. 8) coupled with the more informative subjective prior settings suggested in Allenby et al. (2014). Allenby et al. (2014) propose to adjust the standard weakly informative prior settings to $k + 15$ (from $k + 5$) prior degrees of freedom for the IW-prior (where k denotes the dimension of individual demand parameters) in the standard one-component model, and to set the diagonal elements in the prior scale matrix to 0.5 for constrained coefficients and to 1 for unconstrained coefficients. In addition, the subjective prior information for $\bar{\beta}^*$ is increased to $A_{\mu^*} = .1$ (from .01).

However, as described before, the problem with the standard parameterization is that these more informative subjective settings now apply to both constrained, i.e., to be transformed, and to unconstrained coefficients. While these settings yield much more sensible priors for constrained coefficients, they may be unduly informative for unconstrained coefficients.

Figure 2 compares prior distributions based on $R = 1,000,000$ draws from the positively constrained marginal prior for β_+ (left panel) and the unconstrained marginal prior for β_{uc_1} (right panel).[8] In each panel of Fig. 2 the dashed density in orange is from our proposed marginal-conditional specification. The green dash-dotted density is the corresponding marginal prior from Allenby et al. (2014). The figure illustrates the benefit from our proposed parameterization: While marginal priors for the constrained coefficient in the left panel are essentially identical, the standard parameterization coupled with the more informative settings discussed above imply a much more informative marginal prior for unconstrained coefficients than usual. At first sight, the comparison in the right-panel of Fig. 2 seems to suggest that the standard parameterization coupled with the more informative settings from above simply imply less heterogeneity in β_{uc_1} a priori. However, it is important to realize that the increase in prior degrees of freedom in the IW prior will similarly fail to accommodate much more homogenous markets than what is implied by the prior

[8] Without loss of generality, we fix Γ to zero here, see the discussion following Eq. 11.

settings. In fact, it is the joint possibility of extremely homogenous and extremely heterogenous markets under our suggested prior that causes the pronounced peak at zero together with the fat, sub-exponential tails in the right panel of Fig. 2.

Next, we illustrate how the difference in subjective priors translates into different posteriors in the typical large N, small T setting.

5.2 Population distribution and data generation

We generate heterogeneous consumer preferences obeying sign and order constraints in Eq. 16 using the following transformation and distribution:

$$\begin{aligned}
\beta^* &= \begin{pmatrix} \beta^*_{+} \\ \beta^*_{++} \\ \beta^*_{p} \\ \beta^*_{uc_1} \\ \beta^*_{uc_2} \end{pmatrix} = g^{-1}(\beta) = \begin{pmatrix} \ln(\beta_{+}) \\ \ln(\beta_{++} - \beta_{+}) \\ \ln(-\beta_p) \\ \beta_{uc_1} \\ \beta_{uc_2} \end{pmatrix} \sim N\left(\bar{\beta}^*, V_{\beta^*}\right), \text{ with :} \\
\bar{\beta}^* &= \begin{pmatrix} 0.5 & -0.5 & 0.8 & 2.5 & 2.5 \end{pmatrix}' \text{ and} \\
V_{\beta^*} &= \begin{pmatrix} 0.4 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.2 & -0.15 & 0 & 0 \\ 0 & -0.15 & 0.4 & -0.05 & 0.05 \\ 0 & 0 & -0.05 & 2 & 0 \\ 0 & 0 & 0.05 & 0 & 4 \end{pmatrix}
\end{aligned} \tag{17}$$

Table 2 summarizes the marginal distributions of data generating preferences in the population. Consumers have a decent preference for the two levels of $A1$ and are relatively price sensitive on average. Preferences for the two levels of $A2$ have the same expected value, but are more heterogeneous for the second level. Preferences for the first and second level of $A1$ correlate positively. Furthermore, consumers who prefer the second level of $A1$ are less price sensitive on average, $Cov(\beta^*_{++}, \beta^*_p) = -0.15$. Similarly, consumers who prefer the first level of $A2$ are less price sensitive while preferences for the second level correlate positively with the absolute value of the price coefficient.

We generate a sample of $N = 1000$ consumers with preferences $\{\beta_i\}$ from this population distribution as input to generating discrete choice data Y. Each choice is from the full set of product alternatives at different, randomly drawn prices from a uniform distribution with support in [0.5, 3], plus an outside good. Consequently, there are $p = 5$ alternatives in each choice set. We fix the amount of individual level information at $T = 4$. Recall that many discrete choice studies in marketing barely

Table 2 Summary of marginal distributions of data generating coefficients

	β_{+}	β_{++}	β_p	β_{uc_1}	β_{uc_2}
Median	1.65	2.31	-2.22	2.50	2.50
Mean	2.00	2.68	-2.71	2.50	2.50
Variance	2.00	2.38	3.65	2.00	4.00

Table 3 Mapping between data generating and estimated (identified) parameters illustrated in one choice set

Alternative	Data generating utility	Estimated utility
$(A1_{L1}, A2_{L1}, P_1)$	$(\beta_+ + \beta_{uc_1}) + P_1\beta_p$	$\beta^{id}_{uc_1} + P_1\beta^{id}_p$
$(A1_{L2}, A2_{L1}, P_2)$	$(\beta_{++} + \beta_{uc_1}) + P_2\beta_p$	$(\beta^{id}_{++} + \beta^{id}_{uc_1}) + P_2\beta^{id}_p$
$(A1_{L1}, A2_{L2}, P_3)$	$(\beta_+ + \beta_{uc_2}) + P_3\beta_p$	$\beta^{id}_{uc_2} + P_3\beta^{id}_p$
$(A1_{L2}, A2_{L2}, P_4)$	$(\beta_{++} + \beta_{uc_2}) + P_4\beta_p$	$(\beta^{id}_{++} + \beta^{id}_{uc_2}) + P_4\beta^{id}_p$
Outside	0	0

reach one choice task per parameter to estimate at the individual level. The sparse individual level data scenario assumed in this simulation is therefore representative of applications in practice.

We remove the column pertaining to the first level of A_1 from the design matrix for identification.[9] Table 3 shows the mapping between data generating and identified parameters derived from the design matrix. Since we delete the first level of $A1$ from the design, it follows that $\beta^{id}_{++} = \beta_{++} - \beta_+$, $\beta^{id}_p = \beta_p$, $\beta^{id}_{uc_1} = \beta_{uc_1} + \beta_+$ as well as $\beta^{id}_{uc_2} = \beta_{uc_2} + \beta_+$.

5.3 Estimates of heterogeneity

Figure 3 illustrates the benefits of our proposed marginal-conditional decomposition of the hierarchical prior distribution (see Eqs. 9, 10 and 11) compared to the standard formulation (see Eq. 8) coupled with informative subjective prior settings (Allenby et al. 2014) using the example of the *unconstrained* coefficients $\beta^{id}_{uc_1}$ and $\beta^{id}_{uc_2}$.[10]

It is visually apparent that the standard parameterization (Eq. 8) that cannot but impose informative priors on *both* constrained *and* unconstrained parameters, when constrained parameters require more informative priors, underestimates the amount of preference heterogeneity in the unconstrained coefficients (see the green dashed-dotted densities in Fig. 3). Note that the bias from unduly informative priors on

[9] In principle, the MCMC sampler could navigate the unidentified model at the individual level based on a proper (hierarchical) prior. Non-identification implies that two different vectors of preferences β^1 and β^2 with $\beta^1 \neq \beta^2$ can achieve the exact same likelihood maximum. In an unidentified model, the sampler then generates from the infinite number of different states of the same (high) likelihood for any individual i. However, this interferes with measuring preference heterogeneity in the population. Consider the case of two brands offered in a choice set without an outside option. Only the relative brand preference is likelihood-identified. Now consider two different individuals i and j having the exact same relative preferences. We could set $\beta_i = \begin{pmatrix} \beta_{i1} - \varepsilon & \beta_{i2} - \varepsilon \end{pmatrix}'$ as well as $\beta_j = \begin{pmatrix} \beta_{j1} + \varepsilon & \beta_{j2} + \varepsilon \end{pmatrix}'$ and create arbitrarily large preference heterogeneity for $\varepsilon \to \infty$, while the likelihood of observed choices remains constant.

[10] Note that the normal hierarchical prior for $\beta^{id}_{uc_1}$ and $\beta^{id}_{uc_2}$ used in estimation no longer exactly corresponds to the data generating heterogeneity distribution in this example. The data generating marginal distributions of $\beta^{id}_{uc_1}$ and $\beta^{id}_{uc_2}$ are sums of normally and log-normally distributed random variables, as per our identification constraint, and a mixture of normals prior may further improve generalizations to the population based on the (posterior of) the hierarchical prior.

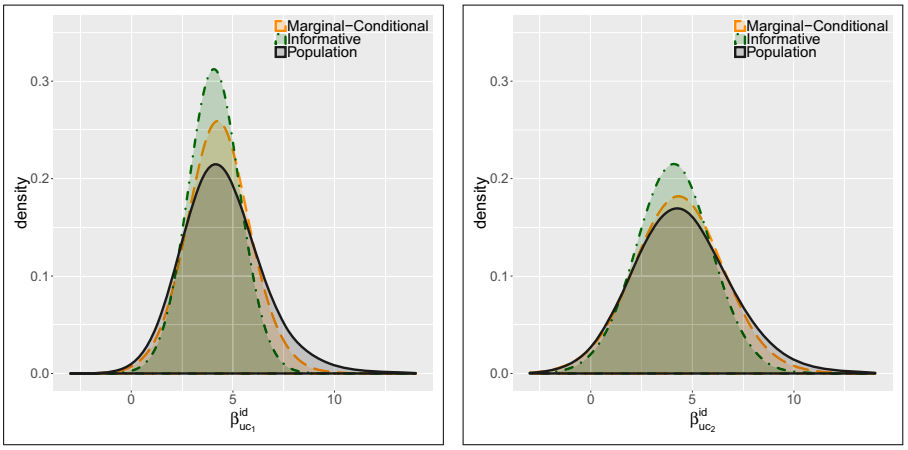

Fig. 3 Posterior predictive population distributions of $\beta^{id}_{uc_1}$ and $\beta^{id}_{uc_2}$ using the marginal-conditional decomposition and the standard formulation ($T = 4$)

unconstrained coefficients further amplifies in the context of a mixture of normals prior where fewer observational units contribute likelihood information about the amount of heterogeneity in each mixture component (see Section 6). Finally, Appendix A.5 reports MH-acceptance rates and MCMC trace plots for a qualitative gauge of the numerical performance of the proposed MCMC algorithm that relies on the marginal-conditional decomposition of parameters in the hierarchical prior.

6 Preferences for fresh hen's eggs

Our first empirical application analyzes Nielsen data on purchases of fresh hen's eggs by German households (see Kotschedoff and Pachali 2020). It illustrates the empirical relevance of the proposed marginal-conditional decomposition of the hierarchical prior. In Germany, eggs are differentiated in terms of animal welfare as summarized in Table 4.

Table 4 Main differences between egg breeding categories

Egg label	Hens per m^2	Surface per hen in cm^2	Outdoor area per hen in m^2	Additional points
Organic	6	1667	4	Organic feed, no beak trimming, no regular use of antibiotics
Free-range	9	1100	4	Live in open barns
Barn	9	1100	0	Live in open barns
Battery	18	550	0	Live in cages

Source: http://www.deutsche-eier.info/die-henne/haltungsformen/; accessed 2 March 2016.

Since 2004, EU regulations require labeling the breeding category on egg packages and printing a code on each single egg indicating origin and breeding category. Consumers associate the four breeding categories with different quality levels: battery eggs $\precsim$ barn eggs $\precsim$ free-range eggs $\precsim$ organic eggs. In 1999 the EU decided that all member states ban the production of battery eggs by 2012. Germany implemented this ban already in 2010. Kotschedoff and Pachali (2020) (KP) use this policy change to evaluate the effect of this increase in minimum quality standard on consumer welfare. They use a sample of 6,961 households who purchased eggs at least four times in the period of 2008 to 2012.[11]

The demand model in KP assumes that households have full information about the egg products offered by the ten retail chains included in the sample. Accordingly, household i's indirect utility from egg product g in chain l at period t is

$$U_{iglt} = \gamma_{i,g} + \alpha_i p_{glt} + \beta_i \mathbf{1}\{units_g = 6\} + \psi_{i,l} + \varepsilon_{iglt}, \tag{18}$$

where $g \in \{Battery, Barn, Free\text{-}range, Organic\}$ and $l \in \{1, \ldots, 10\}$. The indicator variable, $\mathbf{1}\{\}$, denotes whether egg label g has the package size six instead of ten eggs. The price is given by p_{glt} and the mean utility of the outside option is normalized to zero, $u_{iglt} = 0$. The error terms ε_{iglt} is assumed to follow a type I extreme value distribution, as standard in the literature.

KP state that flexible estimation of the retail chain preference coefficients $\{\psi_{i,l}\}$ is particularly important in their demand specification, alleviating a potential bias from the full information assumption implicit to Eq. 18: It is crucial that retail chain preference coefficients become very negative—potentially approaching negative infinity—for those chains a household never or very infrequently purchased eggs from. If a retail chain is estimated to be extremely unattractive to a consumer, the egg prices charged at this chain will not affect this consumer's egg purchasing decisions, independent of the consumer's actual price knowledge set. In addition, KP rely on the inferred information about $\{\psi_{i,l}\}$ when modeling competition among retail chains in a supply side model.

Here, we rely on the simplified demand framework in Eq. 18 to illustrate the benefits of our marginal-conditional decomposition model as developed in Section 3.[12] The model is an example of the typical application featuring a mix of constrained and unconstrained coefficients in the context of a hierarchical model. While we cannot a priori constrain preferences for the retail chains and the battery egg taste coefficient, which measures preferences for battery eggs over the outside good, it seems meaningful and actually important to constrain the remaining parameters. This is because the amount of price variation across quality tiers in this data vastly exceeds the amount of temporal price variation within quality tiers. As a consequence, a household who

[11] Furthermore, they only consider purchases at the top ten retail chains and define boiled and painted eggs as well as eggs from other type of poultry, e.g., quails and gooses, as outside good.

[12] KP in addition control for seasonality and regime changes. However, these controls are irrelevant for the purpose of the illustration here.

Table 5 Restricted attributes and constraints imposed on levels

Restricted Attributes	Constraints
Price	$\alpha \leq 0$
Package size	$\beta \leq 0$
Egg label	$\gamma_{Battery} \leq \gamma_{Barn} \leq \gamma_{Free\text{-}range} \leq \gamma_{Organic}$

is only observed to purchase the highest price alternative (organic eggs) could be rationalized as exhibiting positive preferences for high prices in a model without economically motivated constraints. Similarly, an unconstrained model could misleadingly rationalize the choice pattern of a household who only purchased the lowest price alternative (battery eggs) based on higher (direct utility) preferences for battery eggs than for qualitatively superior alternatives.

We thus employ the constraints summarized in Table 5. Preferences for the four different egg labels should satisfy the quality ordering implied by Table 4 to identify the price coefficient. Everything else equal, for example, a household should not be worse off consuming an organic egg instead of a battery egg. Furthermore, the coefficient for the smaller package size and the price coefficient are constrained to be negative.

Table 6 provides an overview of the number of egg purchase incidents across households in the estimation sample. For most households, we observe a decent number of purchases, resulting in "positive degrees of freedom" at the individual level. The lack of individual level information motivating the use of a hierarchical model is due to the small amount of within quality tier price variation as compared to price variation across quality tiers.

We compare our model (see Eqs. 9 to 11) to the standard formulation (see Eq. 8) coupled with the informative subjective prior advanced in Allenby et al. (2014). These authors propose a somewhat tighter IW-prior for the variance-covariance matrix in a three component mixture of multivariate normals with prior degrees of freedom equal to $k + 25$ (where k is the dimensionality of the individual level model). In addition, they set the diagonal elements in the prior scale matrix to 0.5 for unconstrained coefficients and to 0.05 for constrained coefficients in each normal component. Note that we adjust prior degrees of freedom to $k + 40$ accounting for the fact that, similar as in KP, we rely on a five component (instead of a three component) mixture of normals model in the estimation below.

Compared to the informative subjective prior advanced in Allenby et al. (2014), our marginal-conditional decomposition of the hierarchical prior distribution enables the analyst to be differentially informative about the distribution of constrained and

Table 6 Distribution of the number of egg purchase incidents across $N = 498$ households used in the estimation sample

	Min.	1st Qu.	Median	Mean	3rd. Qu.	Max.
Purchases	4	21	45	56	81	283

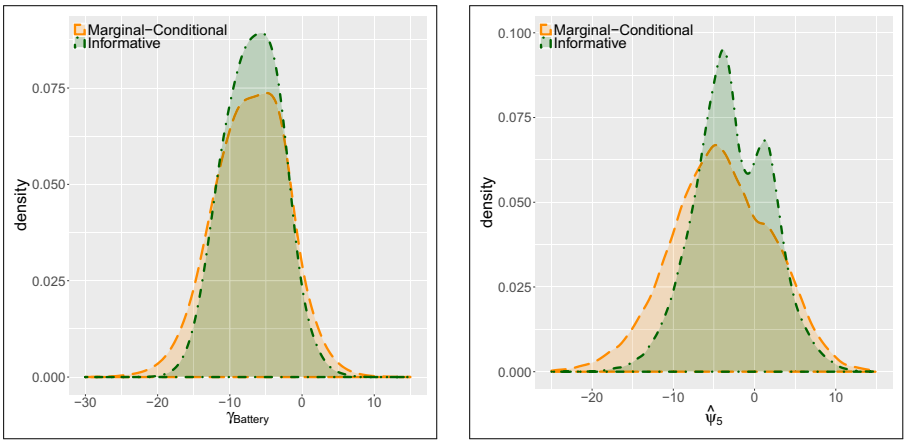

Fig. 4 Posterior predictive population distributions using the marginal-conditional decomposition model and a standard model with informative priors for the battery egg coefficient (left panel) as well as the preference coefficient of the fifth retail chain (right panel)

unconstrained parameters in the population a priori. We use the following subjective settings affecting the prior of constrained coefficients in every mixture component: $\bar{\mu}_c^* = \begin{pmatrix} 0 & \dots & 0 \end{pmatrix}'$, $A_{\mu_c^*} = 0.1 I_{k_c}$, $\bar{V}^* = 0.05 \nu_{V^*} I_{k_c}$ as well as $\nu_{V^*} = k_c + 40$, where I_{k_c} is the identity matrix of dimension $k_c \times k_c$. However, in contrast to Allenby et al. (2014), we can elect to use standard weakly informative, "barely proper" priors for parameters in the conditional prior of unconstrained coefficients: $\bar{\gamma}_z = \begin{pmatrix} 0 & \dots & 0 \end{pmatrix}'$, $A_{\Gamma_z} = 0.01 I_{(k_c+1)}$, $\bar{\Sigma} = \nu_\Sigma I_{k_{uc}}$ as well as $\nu_\Sigma = k_{uc} + 5$.

Table 7 Variances of marginal posterior densities unconstrained preference coefficients implied by the marginal-conditional decomposition model and a standard model with informative priors

	Marginal-Conditional	Informative
Battery	25.0	15.8
Chain 2	60.2	36.1
Chain 3	19.0	13.7
Chain 4	42.7	19.9
Chain 5	38.8	22.0
Chain 6	21.1	14.9
Chain 7	21.8	13.4
Chain 8	25.6	15.6
Chain 9	24.1	15.6
Chain 10	38.3	15.9

Table 8 Comparison of log marginal likelihood values across model specifications

Marginal-Conditional parameterization	Standard parameterization
(selectively informative)	(informative)
−38853	−39001

In order to reduce implied computation times, we draw a random subsample of $N = 498$ households and estimate a model with a five components mixture of normals prior under these two different subjective prior settings.[13]

Figure 4 shows posterior predictive population distributions for the (unconstrained) battery egg coefficient as well as the coefficient measuring preferences for retail chain 5.[14] Both graphs in Figure 4 confirm the finding from the simulation study in Section 5: By imposing an informative prior on all coefficients (that is really needed for the constrained coefficients only) the standard formulation results in the dashed-dotted densities in green, which underestimate heterogeneity in these unconstrained coefficients. This is particularly apparent in the right panel of Fig. 4, where the marginal posterior from the standard parameterization of the hierarchical prior (see Equation 8)—when coupled with informative subjective priors needed to "discipline" the distribution of constrained coefficients—fails to accommodate extremely negative preferences for retail chain 5 in the left tail.

Table 7 summarizes variances of marginal posterior predictive densities of unconstrained coefficients and verifies that the differences across the two subjective prior specifications are substantial. Finally, Table 8 compares model fit based on the Newton-Raftery estimator of the log marginal likelihood. As one may expect, the indistinctively informative specification in the standard prior parameterization (see Eq. 8) translates into inferior fit compared to the informative specification that selectively targets constrained coefficients facilitated by the marginal-conditional decomposition in Eqs. 9 to 11.

7 Tablet PC preferences

Our second empirical application uses data from a commercial discrete-choice conjoint study investigating demand for tablet PCs ("tablets"). Here, we focus on the drawbacks of relying on individual level posterior means $\hat{\beta}_i = \int \beta_i \, p(\beta_i|Y, y_i) d\beta_i$ for market simulation (as defined in Section 2), and estimate implied losses in profits when relying on this method for decision-making. For estimation, we rely on the marginal-conditional decomposition of the hierarchical prior (see Section 3). We show how using posterior means translates into systematic over estimation of

[13] We run both MCMC samplers for $R = 120,000$ iterations and keep every 40th draw. We then burn-off the first 2000 draws and perform our analysis based on the remaining 1000 draws from the converged posterior distribution. We assess convergence by inspecting time-series plots of draws, both at the level of individual respondents and in the hierarchical prior.

[14] We estimate individual retail chain preferences relative to a baseline chain for likelihood identification, i.e., for $l \neq 1$, $\hat{\psi}_{i,l} = \psi_{i,l} - \psi_{i,1}$ measures household i's preference for the lth retailer relative to the first as the baseline level.

Table 9 Attributes and levels in the tablet experiment

Attributes	Levels
Resolution (RE)	Standard (S), High (H)
Memory	8GB, 16GB, 32GB, 64GB, 128GB
SD-Slot	With (SD), Without (SD^-)
Performance (PER)	1 GHz (S), 1.6 GHz (H), 2.2 GHz (VH)
Battery run time (RUN)	4-8 hours (S) , 8-12 hours (H)
Connections (CO)	WLAN (S), WLAN + UMTS (3G), WLAN + LTE (4G)
Synchronization to smartphone	No (SYN^-), Yes (SYN)
Value pack	No (VP^-), Yes (VP)
Equipment	No (EQ^-) , Cover (C), Keyboard (K), Mouse (M), Pencil (P), 32GB Memory Card (32MC), Keyboard & Pencil (KP), Keyboard & Mouse & Pencil (KMP)
Price (P)	Continuous in [99 €,899 €]
Cash back	No (CB^-), 50€, 100€, 150€
Brand (B)	A, B, C, D, E, F, G
Operating system (OS)	A, B
Display size (DS)	7, 8, 10, 12, 13

preferences for sign- and order-constrained attribute levels. Finally, we show empirically how relying on individual level posterior means reduces sign and order violations, in the absence of a theoretically constrained hierarchical prior—arguably a major reason for the popularity of this approach in practice.

Table 9 lists the tablet attributes and attribute levels included in this study. Overall, there are fourteen attributes including a seven level brand attribute. Because of the commercial origin of the data, brand names are disguised. A total of $N = 1046$ respondents participated in this study.

Each respondent evaluated thirteen choice sets ($T = 13$), indicating which if any of the tablets offered in a choice set the respondent would purchase. Each choice set featured three tablets, and an unspecified outside option. Respondents selected the outside or no-buy option in about a quarter (26.6%) of the observed $1,046 \times 13 = 13,598$ choices. Thus, this is a representative example of the type of high-dimensional "large N, small T" studies that have become the standard in industry applications.

The original goal of this study was to help optimize brand A's product design given a fixed set of competitor offerings. As typical of industry grade discrete-choice conjoint studies, the number of parameters at the individual level (36 coefficients after imposing identification constraints) by far exceeds the number of individual level observations. As a consequence, a hierarchical model is required, the hierarchical prior's specification becomes critically important, and—in the likely scenario of heterogeneous preferences—individual level posterior distributions will reflect large amounts of posterior uncertainty about a specific respondent's preferences.

In combination with the ordinal nature of many of the attributes in this study, a standard hierarchical prior specification leads to questionable results. For instance, Fig. 11 and Table 17 in Appendix A.6 showcase that posterior predictive distributions

from an unconstrained hierarchical prior specification coupled with weakly informative subjective priors (e.g., Rossi et al. 2005) clearly violate basic economic intuition. Inferred preferences for levels of cash back refer to the amount of money a customer receives after purchase upon submitting the sales receipt to the manufacturer. According to Table 17 (Appendix A.6), more than 25% of draws from the posterior of the hierarchical prior imply that consumers dislike tablets with larger amounts of cash back. Perhaps even more problematic, the posterior of the hierarchical prior suggests that consumers in the market prefer a tablet with 100€ cash back over the same tablet with 150€ cash back (as indicated by the stochastic dominance of 100€ cash back across all quantiles of the marginal posterior predictive distribution). In a market simulation, this could give rise to the odd outcome that tablets with smaller levels of cash back will be offered at higher prices, everything else equal. Finally, Table 18 (Appendix A.6) shows that the collection of individual level posterior means cuts the support for negative preferences for e.g., 50€ cash back by about 50%. Recall that what may appear as a benefit here is the consequence of measuring heterogeneity inconsistently. These observations call for a diligently constrained hierarchical prior distribution of heterogeneity in the population.

The majority of attributes and levels in Table 9 are such that one can expect every respondent to strictly prefer one level over another level, everything else equal. Table 10 collects all ordinal and sign constraints we thus impose in the hierarchical prior distribution, based on (direct) utility considerations. We constrain preferences for eleven out of the fourteen attributes. We do not impose constraints on brand, operating system, and display size. Although some brands may be preferred on average, it would be wrong to impose the average preference ordering for every respondent, similar with operating systems. Display size may appear as an ordinal attribute at first, but is not once the inconvenience of larger displays in some usage situations,

Table 10 Restricted attributes and constraints imposed on levels

Restricted Attributes	Constraints
Resolution	$\beta_{\text{RE}_H} \geq \beta_{\text{RE}_S}$
Memory	$\beta_{128\text{GB}} \geq \beta_{64\text{GB}} \geq \beta_{32\text{GB}} \geq \beta_{16\text{GB}} \geq \beta_{8\text{GB}}$
SD-Slot	$\beta_{\text{SD}} \geq \beta_{\text{SD}^-}$
Performance	$\beta_{\text{PER}_{VH}} \geq \beta_{\text{PER}_H} \geq \beta_{\text{PER}_S}$
Battery run time	$\beta_{\text{RUN}_H} \geq \beta_{\text{RUN}_S}$
Connections	$\beta_{\text{CO}_{4G}} \geq \beta_{\text{CO}_{3G}} \geq \beta_{\text{CO}_S}$
Synchronization to smartphone	$\beta_{\text{SYN}} \geq \beta_{\text{YSYN}^-}$
Value pack	$\beta_{\text{VP}} \geq \beta_{\text{VP}^-}$
Equipment	$\beta_{\text{EQ}_C}, \beta_{\text{EQ}_K}, \beta_{\text{EQ}_M}, \beta_{\text{EQ}_P}, \beta_{\text{EQ}_{32MC}} \geq \beta_{\text{EQ}^-}$
	$\beta_{\text{EQ}_{KP}} \geq \beta_{\text{EQ}_K}, \beta_{\text{EQ}_P}$
	$\beta_{\text{EQ}_{KMP}} \geq \beta_{\text{EQ}_{KP}}, \beta_{\text{EQ}_M}$
Price	$\beta_{\text{P}} \leq 0$
Cash back	$\beta_{\text{CB}_{150}} \geq \beta_{\text{CB}_{100}} \geq \beta_{\text{CB}_{50}} \geq \beta_{\text{CB}^-}$

or when transporting the tablet, are taken into account. As a consequence, we face a mix of constrained and unconstrained coefficients that we argue is characteristic of most applications of hierarchical models, at least in marketing and economics. We leverage the marginal-conditional decomposition of the hierarchical prior distribution developed in Section 3 to specify suitable subjective priors.

We run the MCMC sampler using the tuned random walk proposal from Section 3 for $R = 500,000$ iterations and keep every 50th draw. We then burn-off the first 8000 draws and perform our analysis based on the remaining 2000 draws from the converged posterior distribution. We assess convergence by inspecting time-series plots of draws, both at the level of individual respondents and in the hierarchical prior. Here, we only report results for a model with a fully parametric, one-component hierarchical prior.[15]

Figure 5 visually compares the marginal posterior predictive population densities of coefficients measuring preferences for levels of the cash back attribute.[16] The utility of the level 'no cash back' is normalized to zero for identification, and individual preferences for 50€, 100€, and 150€ cash back are obtained as $\beta_{\text{CB}_{50},i} = exp(\beta^*_{\text{CB}_{50},i})$, $\beta_{\text{CB}_{100},i} = \beta_{\text{CB}_{50},i} + exp(\beta^*_{\text{CB}_{100},i})$, and $\beta_{\text{CB}_{150},i} = \beta_{\text{CB}_{100},i} + exp(\beta^*_{\text{CB}_{150},i})$, respectively. This way, the coefficient measuring the preference for 50€ relative to no cash back is constrained to be positive, and coefficients associated with more cash back are constrained to be weakly larger than those associated with less cash back.

The upper left panel of Fig. 5 shows inferred population preference distributions for 50€ cash back relative to no cash back (the dash-dotted density in red). Now, if one imposes the constraints we use here and characterizes population preferences using individual level posterior means, the dashed blue density results. Because of the skewness of population preferences as a function of ordinal preferences, individual level posterior means now measure both mean preferences and heterogeneity in the population inconsistently. The mode is biased in the direction of the distribution's skewness, i.e., in the direction of stronger preferences for 50€ cash back relative to the baseline. Compared to the population distribution implied by the posterior of the hierarchical prior, relying on the collection of individual level posterior means clearly underestimates the percentage of consumers with only weak preferences for 50€ cash back. The remaining two panels show how this bias persists, if not accentuates for 100€ and 150€ cash back.

Figure 6 illustrates inferred population preference distributions for display size 8 and 10. We see—in line with the illustration in Appendix A.1—that the collection of individual level posterior means underestimates the degree of taste heterogeneity for these two display sizes.

[15] We find that adding more normal components in a semi-parametric mixture model does not improve holdout predictions.

[16] The marginal posterior of the hierarchical prior and lower level model (n.s.) predictive distributions are essentially identical in this application. Hence we focus on the former.

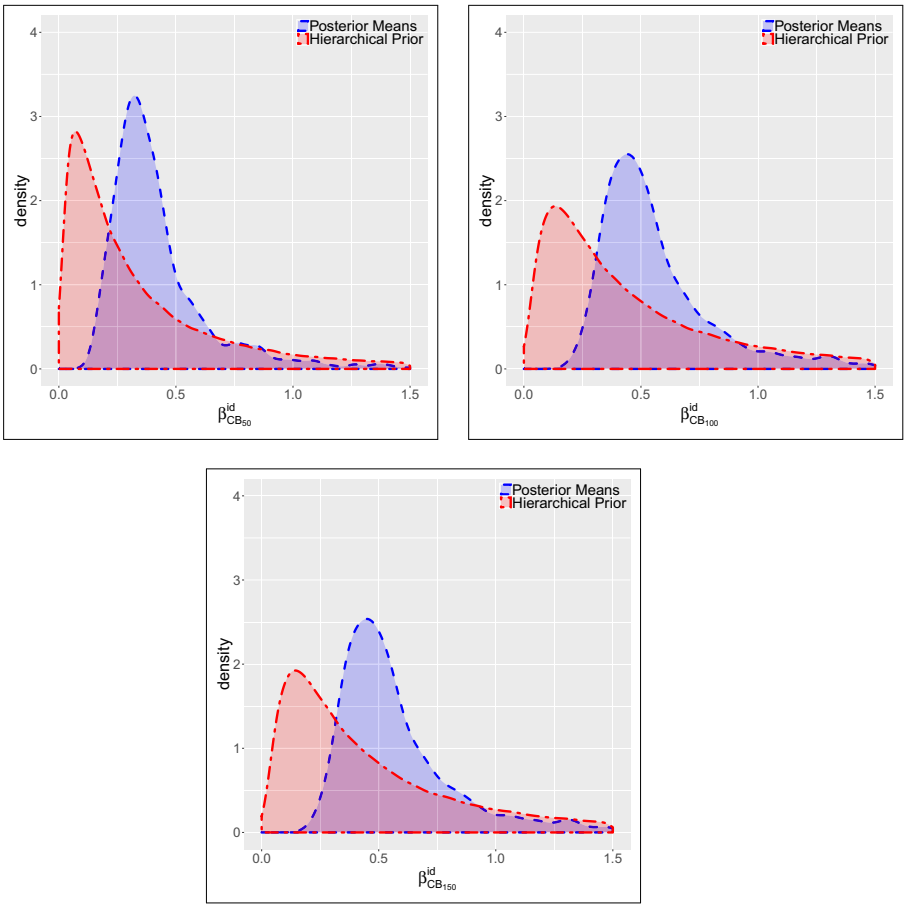

Fig. 5 Posterior predictive population densities for the levels of the cash back attribute using posterior means and the posterior of the hierarchical prior

7.1 Predictive Performance and losses in profits

Next we illustrate the implications of these biases for predictive performance. We use the holdout log-likelihood (HLL) as a measure of how well the two forms of generalization predict choices of holdout respondents, i.e., individuals that were not part of the estimation sample. While it is common to report hit probabilities and hit rates, holdout log-likelihoods are the adequate measure if the eventual target is the prediction of market shares. The holdout likelihood (HL) of individual $h \in \{1, \ldots, H\}$ is defined as the probability of observing the choices $y_h \in Y_{\text{hold}}$ implied by the model after fitting it to the training data Y_{train}. When relying on the posterior of the hierarchical prior and the collection of individual level posterior means, the HL of

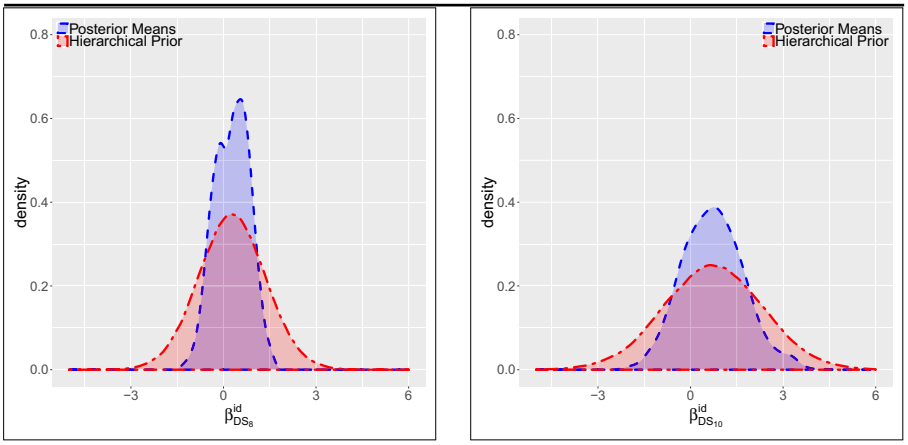

Fig. 6 Posterior predictive population densities of display size 8 (left panel) and 10 (right panel) coefficients using posterior means and the posterior of the hierarchical prior

individual h's choices is defined as in Eqs. 19 and 20 , respectively. In each case $HLL(Y_{\text{hold}}) = \sum_{h=1}^{H} ln(HL(y_h))$.

$$HL(y_h) = \int MNL\left(y_h|g(\beta_h^*)\right) p\left(\beta_h^*|\bar{\beta}^*, V_{\beta^*}\right) p\left(\bar{\beta}^*, V_{\beta^*}|Y_{\text{train}}\right) d\left(\beta_h^*, (\bar{\beta}^*, V_{\beta^*})\right) \tag{19}$$

$$HL(y_h) = \frac{1}{N}\sum_{i=1}^{N} MNL\left(y_h|\hat{\beta}_i\right), \tag{20}$$

We evaluate the predictive performance of the population preference distributions inferred from the collection of individual level posterior means and the posterior of the hierarchical prior using five-fold cross validation. K-fold cross-validation is a common approach to compare the predictive performance of different models for model choice (see e.g., Bishop 2006). We split the complete set of $N = 1046$ choice vectors randomly into five disjoint subsets of approximately the same size. Y^k_{train} and Y^k_{hold} denote the k-th training and holdout sample, containing the data from about 800

Table 11 Predictive performance (holdout log-likelihoods, five-fold cross-validation) of different forms of generalization

	Posterior Means	Hierarchical Prior
Fold 1	−2499	−2466
Fold 2	−2606	−2598
Fold 3	−2810	−2755
Fold 4	−2761	−2718
Fold 5	−3512	−3454
Mean	−2837	−2798

Table 12 Specification of products offered by brand A's competitors

	RE	ME	SD	PER	RUN	CO	SYN	VP	EQ	P	CB	OS	DS
Brand C	High	16GB	Without	1.6 GHz	8-12 h.	4G	Yes	Yes	K	650€	50€	B	10
Brand D	Standard	64GB	With	2.2 GHz	4-8 h.	4G	No	No	No	499€	No	A	10
Brand G	High	32GB	Without	1 GHz	8-12 h.	4G	No	Yes	KP	799€	150€	A	12

(4 folds) and 200 (1 fold) respondents, respectively. The cross-validation estimator for the holdout log-likelihood is defined as the average of the holdout log-likelihoods across the five disjoint holdout data sets (Bengio and Grandvalet 2004):

$$\begin{aligned} \mathrm{CV}_{\mathrm{HLL}}(Y) &= \frac{1}{K}\sum_{k=1}^{K}\sum_{y_h \in Y^k_{\mathrm{hold}}} \mathrm{HLL}\left(A(Y^k_{\mathrm{train}}), y_h\right) \\ &= \frac{1}{K}\sum_{k=1}^{K} \mathrm{HLL}\left(A(Y^k_{\mathrm{train}}), Y^k_{\mathrm{hold}}\right), \end{aligned} \tag{21}$$

$\mathrm{HLL}(A(Y^k_{\mathrm{train}}), y_h)$ denotes the predictive log-likelihood for holdout individual h in the k-th fold computed conditional on training data Y^k_{train} as input (see Eqs. 19–20). The computations always use the same hierarchical Bayes model re-estimated using the respective training data, but summarized either using the collection of individual level posterior means, or the posterior of the hierarchical prior.

Table 11 summarizes the cross-validation results. A random guess for the choices of holdout respondents results in an average log-likelihood of −3770 across our five folds of data. Thus, the hierarchical model yields a decent improvement relative to random predictions, regardless of how the model is summarized for predictions to choices by new respondents. In terms of the comparison between relying on the collection of individual level posterior means and the posterior of the hierarchical prior, the latter outperforms the former not only on average but also in every single fold.[17]

Next we investigate the optimal product configuration for brand A. There are 460,800 product opportunities for brand A in this study. We assume that brand A a priori fixes the levels of some attributes in order to make this problem manageable in the context of varying cost scenarios. We assume that brand A only offers tablets with operating system A, 8-inch display, no SD slot, a 32GB memory card, no smartphone synchronization, and 50€ cash back. These assumptions reduce the action space to 360 unique product possibilities. For a market scenario, we assume that brands C, D, G are already in the market (Table 12).

[17]The five-fold cross-validation log-likelihoods using the unconstrained model are −2997 and −2894 based on posterior means and the posterior of the hierarchical prior, respectively. Constraining the hierarchical prior therefore improves the predictive performance of the model, and regardless of how the model is translated into posterior predictions.

Table 13 Minimum, mean and maximum of product-specific costs illustrated for five cost scenarios

	1st	5th	10th	15th	20th
Min.	30	48	71	93	116
Mean	31	74	127	180	234
Max.	31	101	189	276	364

To more generally capture differences between optimal actions implied by the different approaches of generalizing to the market, we specify a grid of possible costs. This grid comprises 20 different cost settings and is constructed as follows. First, costs are assumed to be the same for the weakest level of each attribute within each scenario. Within attributes, we assume that the cost difference between the baseline and (weakly) preferred levels is determined by a constant factor, i.e. $c_{L2} = f * c_{L1},\ c_{L3} = f * 2 * c_{L1},\ c_{L4} = f * 3 * c_{L1},\ \ldots$, for the levels of a priori ordered attributes; L_1 is the least preferred level. We set $f = 3$ in this example and obtain 20 different scenarios by changing the cost of producing the least preferred levels $\{c_{L1}\}$ of the ordinal attributes to be optimized.

Table 13 summarizes the distribution of product-specific costs across the 360 product opportunities for the first, fifth, tenth, fifteenth and twentieth cost scenario. As can be seen, the grid includes both small as well as large absolute cost differences. In the first cost scenario, it is straightforward for brand A to offer a tablet combining the most attractive attribute levels, i.e., high resolution, 128GB, 2.2 Ghz, $8 - 12$ hours battery, WLAN + LTE (4G), and a value pack, from the attributes to be optimized. As cost differences between attribute levels increase, it becomes less and less profitable to offer this high quality combination of attributes and we compute the expected loss caused by relying on a suboptimal form of generalization each time.

Table 14 summarizes the distribution of brand A's expected percentage losses incurred by relying on the collection of individual level posterior means relative to inferred actions based on the posterior of the hierarchical prior a_{hp} across cost scenarios. We find that optimization results that rely on the collection of individual level posterior means to represent market preferences are clearly inferior and the average percentage loss of 6.68% from using this latter method seems substantial.

Table 14 Percentage losses from using posterior means across cost scenarios relative to optimal actions from the posterior of hierarchical prior

	Minimum	Mean	Maximum
Posterior Means	1.162	6.683	12.193

8 Discussion

Models of consumer heterogeneity play a pivotal role in marketing and economics. Typical applications are random coefficients or mixed logit models for aggregate or panel data and hierarchical Bayesian models. Historically, statistical efficiency or computational arguments motivate the choice of heterogeneity model (e.g., Allenby and Ginter 1995 and Lenk et al. 1996). However, what can be learned about and subsequently extrapolated from the inferred heterogeneity distribution is limited by functional form assumptions such as e.g., the assumption of multivariate normally distributed preferences. For example, consistent estimates of the first and second moments, and correlations in the heterogeneity distribution—all which can be accomplished based on a multivariate normal prior—will fail to translate into useful market simulators in the context of highly non-normal distributions, e.g., distributions that are highly asymmetric.

Various semi-parametric formulations have been advanced (e.g., Lenk and DeSarbo 2000, Li and Ansari 2014 and Rossi 2014) to overcome the often unrealistic assumptions about higher moments inherent to the multivariate normal prior. The additional flexibility afforded by semi-parametric formulations is an important step towards more faithful prior population formulations. However, if as usual the parametric component in a semi-parametric model provides full prior support for all coefficients in a model, the semi-parametric model should still be considered atheoretical and thus mis-specified from an economic point of view. For example, a mixture of normals a priori supports positive price coefficients and this support vanishes a posteriori only in limiting cases of little practical relevance.

The problem with standard, statistically motivated prior population distributions has been recognized in the academic literature early on (see the pioneering contribution by Boatwright et al. 1999), but no general solution has emerged. Recently, Allenby et al. (2014) introduced an informative subjective prior specification for log-normal hierarchical priors. These priors are easily implemented (compared to the truncated normal in Boatwright et al. 1999), but require the analyst to depart from the standard weakly informative subjective prior settings in hierarchical models (e.g., Rossi et al. 2005). In the common situation where the heterogeneity distribution comprises both constrained and unconstrained coefficients (e.g., brand and price coefficients), the choice of subjective prior parameters is an unresolved problem for which this paper proposes a solution.

The contribution of this paper is a marginal-conditional decomposition of the population distribution that allows researchers to be informative about constrained parameters, on a logarithmic scale, while retaining maximal flexibility regarding the (conditional) hierarchical prior of unconstrained coefficients. The suggested specification is easily implemented and the additional computational effort is minimal.

Our specification becomes essential whenever the heterogeneity distribution comprises both constrained and unconstrained coefficients such as e.g., in heterogeneous or mixed choice models that feature brand coefficients and a price coefficient. Finally, we develop how to tune individual level proposal densities for numerically efficient MCMC inference in the presence of sign- and order-constraints. This generalization of pre-tuned proposal densities (Rossi et al. 2005) is particularly important in high dimensional models that feature a multiplicity of constraints.

We thus overcome the choice between a mis-specified heterogeneity distribution and a the common ad-hoc use of the collection of individual level means that fail to measure heterogeneity consistently. The marginal-conditional decomposition developed in this paper facilitates the formulation of more economically faithful heterogeneity distributions based on prior constraints, broadening the applicability of hierarchically formulated choice and demand models in marketing and economics.

An aspect of the subjective prior for order constrained coefficient that we have not explored in this paper, but plan to investigate in future research, is that of prior scale differences and dependence between coefficients for an ordinally constrained attribute. It is easy to verify by simulation that prior scale differences and dependence can be used to express structured beliefs about heterogeneity in ordinal preferences. For example, the population could be heterogeneous in their valuation of a lower level of an ordinal attribute but relatively homogeneous in incremental preferences for the next higher level. Alternatively, the population could exhibit substantial heterogeneity in the incremental valuation of the next higher level. Finally, the *amount of* heterogeneity in the increment could be correlated with the valuation of the lower level, such that low, medium, or high valuations of the lower level co-occur with relatively more heterogeneity in the incremental valuation of the higher level.

Last but not least, it could be interesting to compare (a mixture of) multivariate truncated normal distributions to the log-normal prior formulation used in this paper. The recently proposed exchange algorithm can handle the "double-intractability" due to the intractable normalization of the truncated multivariate normal (Möller et al. 2006; Murray et al. 2006; see Kosyakova et al. (2020) for a recent adaptation of the exchange algorithm in marketing).

Acknowledgments We thank the editor and two anonymous reviewers for their constructive comments and suggestions. Any errors are our own.

AppendixAppendix A

A.1 Illustrating inconsistent inference for heterogeneity when based on posterior means of individual coefficients

We illustrate the drawbacks of relying on individual level means in the prototypical "large N small T" situation. We show that aggregating individual level estimates results in inconsistent market level inferences that lack a bias-variance trade-off justification. Therefore, the collection of individual level posterior means is not a valid non-parametric representation of the distribution of heterogeneity in the population.

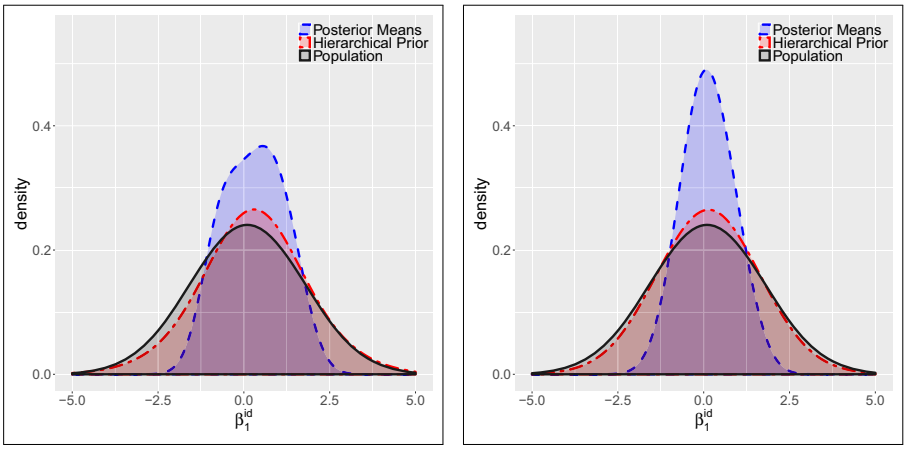

Fig. 7 Posterior predictive population distributions of β_1^{id} from $N = 200$ (left panel) and $N = 3000$ (right panel), $T = 3$

Figure 7 illustrates this aspect in a simplified simulation setting estimating individual level part-worths for five brands.[18] The graphs compare marginal posterior densities of the contrast between the second and the first brand $\beta_2 - \beta_1 = \beta_1^{id}$ in the population for the two sample sizes $N = 200$ and $N = 3000$ in this example. The blue dashed line depicts the distribution of individual level posterior means, the red two-dashed line the distribution implied by the posterior of the hierarchical prior, $p(\bar{\beta}^{id}, V_\beta^{id}|data, prior)$, and the black solid line corresponds to the data generating density.[19] In this example, each consumer provides $T = 3$ choices and each choice set features three randomly chosen brands. Thus, the amount of likelihood information at the individual level is small, reflecting the common situation of "negative degrees of freedom" at the individual level in e.g., choice-based-conjoint analysis (see Lenk and Orme 2009 for a discussion of the trend towards complex individual level models). Comparing posterior densities in the graphs, it is visually apparent that the collection of individual level posterior means—where each individual posterior mean is shrunk towards the population average—results in biased inference about the heterogeneity distribution in the population, and regardless of the number of consumers in the sample. In fact, this bias towards the center of the population preference distribution is *increasing* in the sample size N illustrating how the collection of individual level posterior means as a representation of preference heterogeneity fails consistency in N.

[18] Data generating part-worths are from a multivariate normal distribution, $\beta \sim N(\bar{\beta}, V_\beta)$ with mean $\bar{\beta} = \begin{pmatrix} 0 & 0.1 & 0.2 & 0.3 & 0.4 \end{pmatrix}'$ and variance-covariance matrix $V_\beta = \text{diag}\begin{pmatrix} 1 & 1.5 & 2 & 2.5 & 3 \end{pmatrix}$ representing the population of consumers.

[19] Densities are estimated using a hierarchical Bayesian MNL model over the identifiable parameters with standard weakly informative subjective prior settings as described in e.g., Rossi et al. (2005).

A.2 Posterior distributions: log-normal prior

We set $B_z^{*c} := \left(\iota \; B^{*c}\right)$ in what follows. The posteriors associated with the priors in Eq. 12 are (see e.g., Rossi et al. 2005):

$$\begin{aligned}
\Sigma|B^{*uc}, B_z^{*c} &\sim IW\left(\nu_\Sigma + N, \bar{\Sigma} + S_\Sigma\right) \\
\gamma_z|B^{*uc}, B_z^{*c}, \Sigma &\sim N\left(\tilde{\gamma}_z, \Sigma \otimes \left((B_z^{*c})' B_z^{*c} + A_{\Gamma_z}\right)^{-1}\right), \text{ with} \\
\tilde{\gamma}_z &:= vec(\tilde{\Gamma}_z),\ \tilde{\Gamma}_z = \left((B_z^{*c})' B_z^{*c} + A_{\Gamma_z}\right)^{-1}\left((B_z^{*c})' B_z^{*c}\hat{\Gamma}_z + A_{\Gamma_z}\bar{\Gamma}_z\right), \\
\hat{\Gamma}_z &= \left((B_z^{*c})' B_z^{*c}\right)^{-1}(B_z^{*c})' B^{*uc}, \text{ and} \\
S_\Sigma &= \left(B^{*uc} - B_z^{*c}\tilde{\Gamma}_z\right)'\left(B^{*uc} - B_z^{*c}\tilde{\Gamma}_z\right) + \left(\tilde{\Gamma}_z - \bar{\Gamma}_z\right)' A_{\Gamma_z}\left(\tilde{\Gamma}_z - \bar{\Gamma}_z\right) \quad (22)
\end{aligned}$$

$$\begin{aligned}
\mu_c^*|B^{*c}, V^* &\sim N\left(\tilde{\mu}_c^*, \tilde{A}_{\mu_c^*}\right) \\
V^*|B^{*c}, \mu_c^* &\sim IW\left(\tilde{\nu}_{V^*}, \tilde{V}^*\right) \text{ with} \\
\tilde{A}_{\mu_c^*} &= \left(N(V^*)^{-1} + A_{\mu_c^*}\right)^{-1}, \\
\tilde{\mu}_c^* &= \tilde{A}_{\mu_c^*}\left(\left(\iota' \otimes (V^*)^{-1}\right) vec\left((B^{*c})'\right) + A_{\mu_c^*}\bar{\mu}_c^*\right), \\
\tilde{\nu}_{V^*} &= \nu_{V^*} + N \text{ and } \tilde{V}^* = \bar{V}^* + \left(B^{*c} - \iota(\mu_c^*)'\right)'\left(B^{*c} - \iota(\mu_c^*)'\right) \quad (23)
\end{aligned}$$

A.3 Exact Hessian of transformed variates

The Hessian information about β_i^* in individual i's data is defined as

$$H_i^* = \frac{\partial^2 l_i}{\partial \beta_i^* \partial \beta_i^{*\prime}}, \tag{24}$$

where $l_i := MNL(y_i|g(\beta_i^*))$ denotes individual i's likelihood function.

Taking first derivative yields

$$\frac{\partial l_i}{\partial \beta_i^{*\prime}} = \frac{\partial l_i}{\partial g(\beta_i^*)'}\frac{\partial g(\beta_i^*)}{\partial \beta_i^{*\prime}}, \tag{25}$$

according the chain rule. We define $\nabla l_i := \frac{\partial l_i}{\partial g(\beta_i^*)'}$ as a k-dimensional row vector and $J_g := \frac{\partial g(\beta_i^*)}{\partial \beta_i^{*\prime}}$ as the $(k \times k)$-Jacobian matrix. Accordingly, each element $j \in \{1, \ldots, k\}$ in Equation 25 can be expressed as

$$\left[\frac{\partial l_i}{\partial \beta_i^{*\prime}}\right]_j = \left[\bar{l}_i\right]_j := \nabla l_i J_g^j, \tag{26}$$

where J_g^j denotes the jth column of J_g. Hence

$$H_i^* = \left(\frac{\partial [\bar{l}_i]_1}{\partial \beta_i^*} \ \cdots \ \frac{\partial [\bar{l}_i]_j}{\partial \beta_i^*} \ \cdots \ \frac{\partial [\bar{l}_i]_k}{\partial \beta_i^*} \right), \tag{27}$$

with:

$$\frac{\partial \left[\bar{l}_i\right]_j}{\partial \beta_i^*} = J_g' H_i J_g^j + \frac{\partial J_g^j}{\partial \beta_i^*} \nabla l_i' \tag{28}$$

A.4 Illustrating the value of the proposed tuning

Our small illustration only involves choices by one individual, i.e., no unobserved heterogeneity. Inside goods are characterized by one five level, ordinal attribute:

$$\begin{aligned} \beta^* = g^{-1}(\beta) &= \left(\beta_1 \ \ln(\beta_2 - \beta_1) \ \ln(\beta_3 - \beta_2) \ \ln(\beta_4 - \beta_3) \ \ln(\beta_5 - \beta_4) \right)' \\ &= \left(-1 \ 0.2 \ 0.5 \ -0.1 \ -0.5 \right)' \end{aligned} \tag{29}$$

The individual chooses repeatedly ($T = 20$ and $T = 1000$) from choice sets that contain all five possible inside goods and an outside good with utility normalized to zero according to an MNL model. We compare the numerical performance of our tuned MCMC chain to a simpler, more standard tuning with $\beta_i^{*cand} \sim N\left(\beta_i^*, c^2 I\right)$. Our target quantity are numerical standard errors of posterior means denoted $numSE$ from MCMC chains of length 1,000,000 initialized at data generating values. The numerical standard error approximates the variation in posterior means across different, independent same length runs of the MCMC, after convergence. The tuning parameter c^2 in the simpler, more standard proposal density is optimized targeting the average of numerical standard errors across the five parameters on the grid $\left(0.01 \ 0.06 \ 0.11 \ \ldots \ 1.46\right)$. This parameter is set to its default value of $c^2 = 1$ (see Rossi et al. 2005) in our proposed tuning scheme.

Table 15 Numerical efficiency of MCMC, standard versus proposed tuning, $N = 1$, $T = 20$ and $T = 1000$

	$T = 20$		$T = 1000$	
	Standard	Proposed tuning	Standard	Proposed tuning
$numSE_1$	0.0562	0.0168	0.0116	0.0039
$numSE_2$	0.1232	0.0217	0.0383	0.0397
$numSE_3$	0.0876	0.0232	0.0044	0.0018
$numSE_4$	0.0279	0.0138	0.0030	0.0005
$numSE_5$	0.0632	0.0167	0.0037	0.0007

A.5 Numerical properties of marginal-conditional MCMC algorithm (Section 5)

Table 16 Quantiles of rejection rates of individual level parameter updates among 1,000 simulated individuals

1%	25%	50%	75%	99%
0.64	0.70	0.72	0.74	0.78

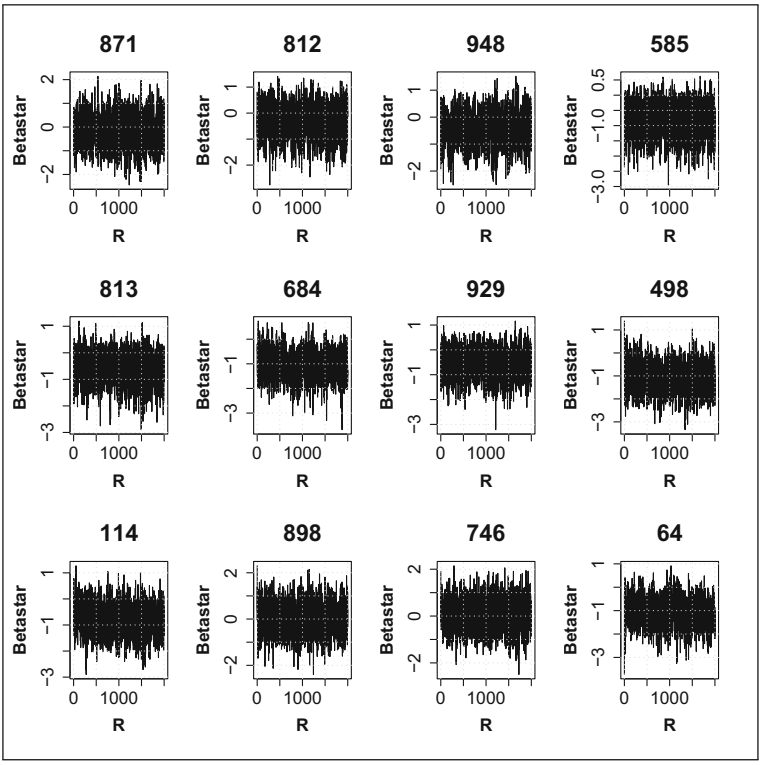

Fig. 8 Individual level coefficients β_+ for 12 randomly chosen consumers

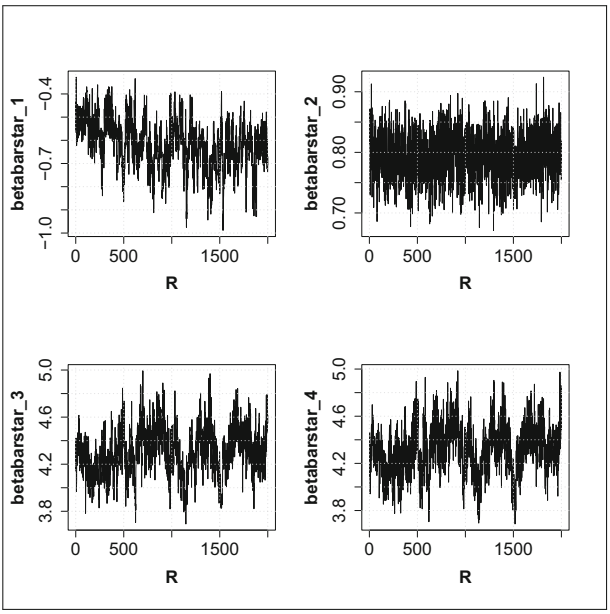

Fig. 9 Upper level mean coefficient $\bar{\beta}^*$: 1 β^*_{+}, 2 β^*_{++}, 3 $\beta^*_{uc_1}$, 4 $\beta^*_{uc_2}$

Figures 8, 9, and 10 show MCMC trace-plots of draws retained for estimation for selected parameters in the simulation study from Section 5.

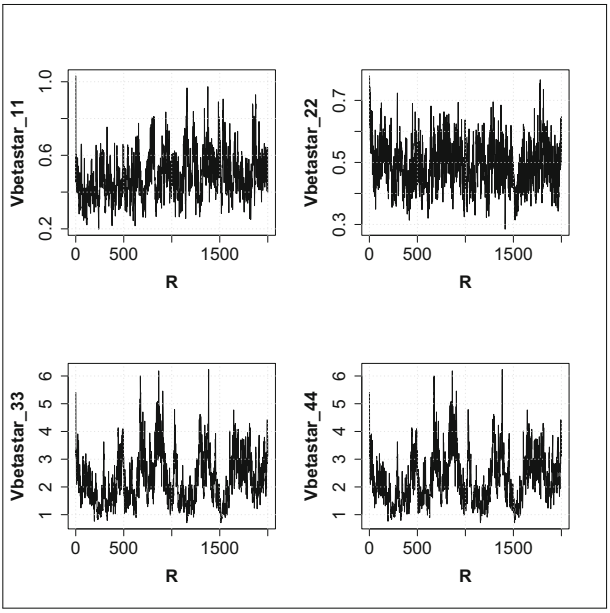

Fig. 10 Upper level variance coefficients V^*_{β}

A.6 Tablet PC preferences in an unconstrained model

Table 17 Quantiles of marginal posterior densities for the levels of the cash back attribute in an unconstrained model

	$\beta^{id}_{CB_{50}}$	$\beta^{id}_{CB_{100}}$	$\beta^{id}_{CB_{150}}$
1%	−2.238	−2.544	−3.317
25%	−0.284	−0.118	−0.655
50%	0.501	0.840	0.371
75%	1.291	1.806	1.399
99%	3.251	4.240	4.060

Table 18 Fraction of sign violations for the 50-cash back attribute level, i.e. $\sum_{r=1}^{R} \mathbb{1}(\beta^{id,r}_{CB_{50}} < 0)/R$, implied by the population distributions based on posterior means and the posterior of the hierarchical prior

Posterior Means	Hierarchical Prior
0.173	0.331

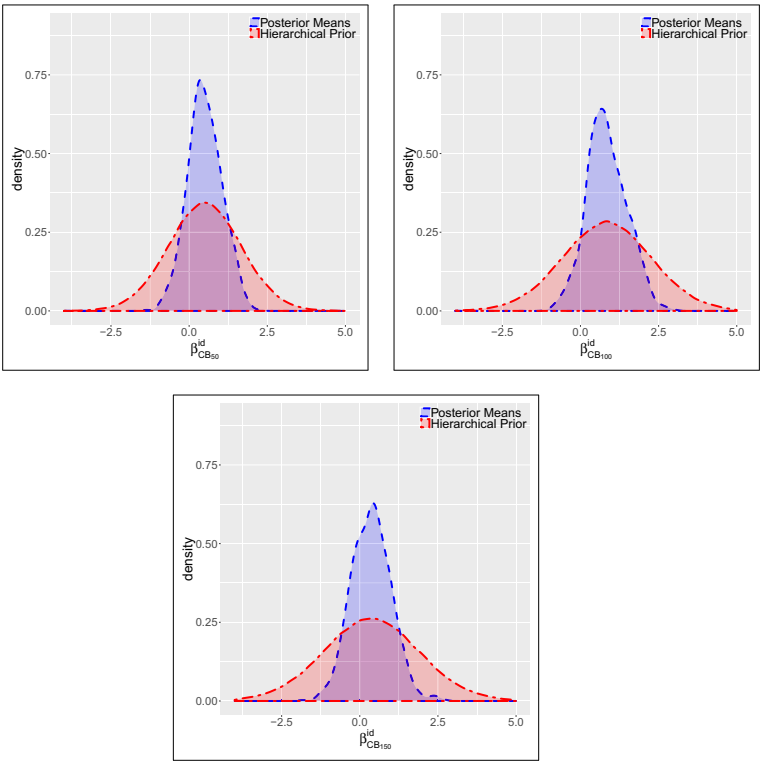

Fig. 11 Posterior predictive population densities of the levels of the cash back attribute using posterior means and the posterior of the hierarchical prior in an unconstrained model

References

Allenby, G.M., Arora, N., Ginter, J.L. (1995). Incorporating prior knowledge into the analysis of conjoint studies. *Journal of Marketing Research*, *32*(2), 152–162.

Allenby, G.M., Arora, N., Ginter, J.L. (1998). On the heterogeneity of demand. *Journal of Marketing Research*, *35*(3), 384–389.

Allenby, G.M., Brazell, J.D., Howell, J.R., Rossi, P.E. (2014). Economic valuation of product features. *Quantitative Marketing and Economics*, *12*(4), 421–456.

Allenby, G.M., & Ginter, J.L. (1995). Using extremes to design products and segment markets. *Journal of Marketing Research*, *32*, 392–403.

Allenby, G.M., & Lenk, P.J. (1994). Modeling household purchase behavior with logistic normal regression. *Journal of the American Statistical Association*, *89*(428), 1218–1231.

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of k-fold cross-validation. *Journal of Mashine Learning Research*, *5*, 1089–1105.

Berry, S., Levinsohn, J., Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, *63*(4), 841–890.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Berlin: Springer.

Boatwright, P., McCulloch, R., Rossi, P.E. (1999). Account-level modeling for trade promotion: An application of a constrained parameter hierarchical model. *Journal of the American Statistical Association*, *94*(448), 1063–1073.

Dubé, J.-P., Hitsch, G.J., Rossi, P.E. (2010). State dependence and alternative explanations for consumer inertia. *The RAND Journal of Economics*, *41*(3), 417–445.

Dubé, J.-P., Hitsch, G.J., Rossi, P.E., Vitorino, M.A. (2008). Category pricing with state-dependent utility. *Marketing Science*, *27*(3), 417–429.

Gelfand, A.E., Smith, A.F., Lee, T.-M. (1992). Bayesian analysis of constrained parameter and truncated data problems using gibbs sampling. *Journal of the American Statistical Association*, *87*(418), 523–532.

Hajivassiliou, V., McFadden, D., Ruud, P. (1996). Simulation of multivariate normal rectangle probabilities and their derivatives theoretical and computational results. *Journal of Econometrics*, *72*(1), 85–134.

Kosyakova, T., Otter, T., Misra, S., Neuerburg, C. (2020). Exact mcmc for choices from menus - measuring substitution and complementarity among menu items. *Marketing Science*, *39*(2), 427–447.

Kotschedoff, M.J.W., & Pachali, M.J. (2020). Higher minimum quality standards and redistributive effects on consumer welfare. *Marketing Science*, *39*(1), 253–280.

Lenk, P.J., & DeSarbo, W.S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93–119.

Lenk, P.J., DeSarbo, W.S., Green, P.E., Young, M.R. (1996). Hierarchical bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, *15*(2), 173–191.

Lenk, P.J., & Orme, B.K. (2009). The value of informative priors in bayesian inference with sparse data. *Journal of Marketing Research*, *46*(6), 832–845.

Li, Y., & Ansari, A. (2014). A bayesian semiparametric approach for endogeneity and heterogeneity in choice models. *Management Science*, *60*(5), 1161–1179.

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

Möller, J., Pettitt, A.N., Reeves, R., Berthelsen, K.K. (2006). An efficient markov chain monte carlo method for distributions with intractable normalising constants. *Biometrika*, *93*(2), 451–458.

Murray, I., Ghahramani, Z., MacKay, D. (2006). Mcmc for doubly-intractable distributions. UAI.

Reiss, P.C., & Wolak, F.A. (2007). Chapter 64 structural econometric modeling: Rationales and examples from industrial organization. Volume 6 of Handbook of Econometrics, pp. 4277–4415. Elsevier.

Revelt, D., & Train, K. (1998). Mixed logit with repeated choices: Households' choices of appliance efficiency level. *Review of economics and statistics*, *80*(4), 647–657.

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

Rossi, P.E., Allenby, G.M., McCulloch, R. (2005). *Bayesian statistics and marketing*. New York: Wiley.

Rossi, P.E., McCulloch, R.E., Allenby, G.M. (1996). The value of purchase history data in target marketing. *Marketing Science*, *15*(4), 321–340.

Sawtooth (2013). The cbc system for choice-based conjoint analysis. Technical paper series, Sawtooth Software.
Train, K.E. (2009). Discrete Choice Methods with Simulation, 2 edn. Cambridge University Press.

Affiliations

Max J. Pachali[1] · Peter Kurz[2] · Thomas Otter[3]

Peter Kurz
p.kurz@bms-net.de

Thomas Otter
otter@marketing.uni-frankfurt.de

1 Tilburg University, Tilburg, Netherlands

2 bms marketing research + strategy GmbH, Munich, Germany

3 Goethe University Frankfurt, Frankfurt, Germany