How to Generalize from a Hierarchical Model?

Max J. Pachali [*] Peter Kurz [†] Thomas Otter [‡]

First version: August 14, 2017
This version: April 12, 2018

Abstract

Thomas: Can you revise the abstract? In many marketing applications of hierarchical models the goal is to inform actions that apply to the population of consumers beyond the sample available for calibration; the goal is to generalize to the population, an exercise often referred to as market simulation. Examples are price and product optimization based on data from discrete choice experiments. It is common practice to rely on the collection of individual level posterior mean preferences of in-sample respondents, or consumers, as a representation of population preferences in this context. We show that this results in biased inferences and misleading recommendations precisely in situations that call for a hierarchical model. Generalizations that avoid this bias rely heavily on the hierarchical prior distribution, which is often only regarded as a smoothing device but not as a useful model per se. We show how to specify more faithful hierarchical prior distributions based on prior constraints and a marginal-conditional decomposition for the hierarchical prior distribution, and how to efficiently sample from the implied posterior. Practical relevance is demonstrated in two illustrative empirical case studies.

Keywords: *discrete choice, Bayesian inference, market simulation, constrained hierarchical prior*

[*]Goethe University, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany; Max.Pachali@hof.uni-frankfurt.de
[†]Kantar TNS, Landsberger Strasse 284, 80687 Munich, Germany; Peter.Kurz@tns-infratest.com
[‡]Goethe University, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt, Germany; otter@marketing.uni-frankfurt.de

1 Introduction

With the advent of modern Bayesian computational methods, applications of hierarchical models have become ubiquitous in marketing, e.g., for making inference from household scanner panel data or from discrete choice experiments (e.g., Allenby, Arora, and Ginter, 1998; Allenby and Lenk, 1994; Dubé, Hitsch, and Rossi, 2010; Rossi, McCulloch, and Allenby, 1996; Sawtooth, 2013). In many applications, the inferential target pertains to a population beyond the sample of consumers or respondents providing the panel data for model calibration. For example, optimal prices or product configurations inferred from the model are expected to be optimal in the population and not just optimal in the observed, finite sample.

From a statistical modeling point of view, the so called "upper level"-model, i.e., the posterior of the hierarchical prior 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 in the hierarchical prior 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. However, only a few papers explicitly rely on the posterior of the hierarchical prior distribution in the model to infer (counterfactual) market outcomes (Allenby, Brazell, Howell, and Rossi, 2014; Ferjani, Jedidi, and Jagpal, 2009; Lenk and Orme, 2009; Rossi et al., 1996; Sonnier, Ainslie, and Otter, 2007).

A likely reason for this is that, historically, statistical efficiency or computational arguments motivate the choice of hierarchical prior (e.g., Allenby and Ginter, 1995; Lenk, DeSarbo, Green, and Young, 1996). Unfortunately, standard hierarchical prior distributions often lack economic rationality. For example, Reiss and Wolak (2007) remark that the estimated distribution of marginal utility of fuel economy in Berry, Levinsohn, and Pakes (1995) suggests that about half of consumers in the car market dislike fuel economy. As another example, Dubé et al. (2010) and Dubé, Hitsch, Rossi, and Vitorino (2008) find posterior support for positive price coefficients in the inferred heterogeneity distribution. Dubé et al. (2010) also note that the posterior support for positive price coefficients essentially vanishes when the heterogeneity distribution is approximated by the collection of individual level posterior means.

We argue that the lack of economic rationality in standard hierarchical prior distributions has materially contributed to the prevailing current practice of using posterior means of individual level coefficients as the basis for market simulation aimed at determining e.g., optimal prices or optimal product configurations (e.g., Elrod, 2001; Huber, Orme, and Miller, 1999).[1] The resulting "spreadsheet-representation" of the distribution of heterogeneity is easy to work with, and substantially reduces sign and order violations implied by the posterior of standard parametric or semi-parametric hierarchical prior distributions (by an order of magnitude in the simulation we report in Section 3). In other words, standard hierarchical priors are generally regarded as highly effective smoothing devices but not as useful models per se.

However a drawback of relying on individual level posterior means in the prototypical "large N small T" situation is that aggregating individual level estimates results in biased market level inferences that lack a bias-variance trade-off justification. Therefore, the collection of individual level posterior means

[1] See also Dubé, Hitsch, and Rossi (2009) who cluster individual level posterior means for a low dimensional discrete representation of heterogeneity as input for equilibrium price calculations in a dynamic model.

are not a valid non-parametric approximation to the distribution of heterogeneity in the population. Figure 1 illustrates this aspect in a simplified simulation setting estimating individual level part-worths for five brands.[2] 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 line depicts the distribution of individual level posterior means, the red line the distribution implied by the posterior of the hierarchical prior, $p(\bar{\beta}^{id}, V_\beta^{id}|data, prior)$; finally, the black line corresponds to the data generating density.[3] 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. While this bias helps to keep sign and order violations at bay, it has implications for forecasting and optimal product choice. For example, optimal products are biased towards the preferences of the average consumer.

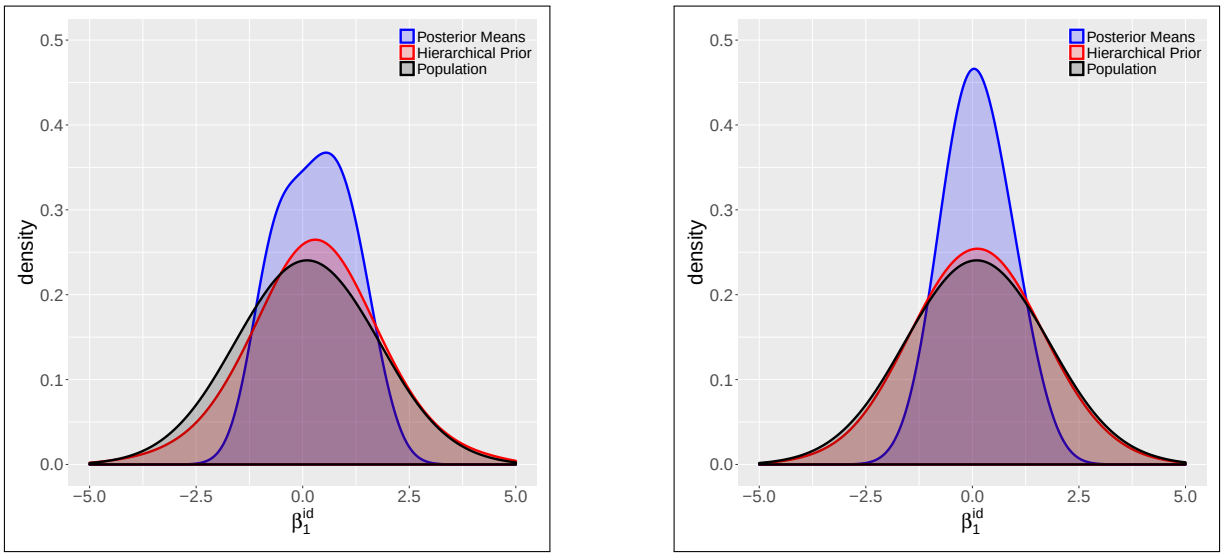

Figure 1: Posterior predictive population distributions of β_1^{id} from $N = 200$ (left panel) and $N = 3000$ (right panel), $T = 3$.

What can be learnt about the hierarchical prior distribution is limited by prior 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

[2]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.

[3]Densities are estimated from a hierarchical Bayesian MNL model over the identifiable parameters with standard weakly informative subjective prior settings as described in e.g., Rossi, Allenby, and McCulloch (2005).

that are highly asymmetric. Various semi-parametric formulations have been advanced (e.g., Lenk and DeSarbo, 2000; Li and Ansari, 2014; Rossi, 2014) to overcome the often unrealistic assumptions about higher moments inherent to the multivariate normal hierarchical prior. The additional flexibility afforded by semi-parametric formulations is an important step towards more faithful hierarchical prior 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 misspecified 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.

While a completely theory driven specification of hierarchical prior distributions appears to be beyond reach, some authors argue in favor of theory driven constraints for the hierarchical prior distribution, see e.g., Boatwright, McCulloch, and Rossi (1999) and Allenby et al. (2014) in the context of fully and semi-parametric hierarchical prior distributions. We build on this idea and develop it further.

In industry grade applications, a prior understanding of preferences in the population often suggest a large number of sign and order restrictions. In the common situation where the heterogeneity distribution thus comprises both constrained and unconstrained coefficients, the choice of sensible prior subjective parameters is an unresolved challenge. We argue that the lack of guidance on how to specify more economically faithful hierarchical prior distributions materially contributes to the popularity of using individual level posterior means for market simulation in practice. The shrinkage of individual level posterior means to a reliably estimated population mean in general reduces the number of sign and order violations, albeit at the expense of biasing inferences about heterogeneity, as shown above.

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 that this decomposition is important whenever the hierarchical prior 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).

In a nutshell the goal of this paper is to facilitate the formulation of more economically faithful hierarchical prior distributions and thereby convince applied academic and industry researchers to abandon market simulators built on posterior means of individual level preferences. The remainder of the paper proceeds as follows: In Section 2 we develop the hierarchical prior formulation and efficient posterior inference using MCMC. Section 3 then investigates the relative performance of the proposed approach using simulated data. Sections 4 and 5 report the results from two empirical illustrations based on data from a discrete-choice experiment on tablet PCs as well as household scanner panel data on fresh hen's eggs (Kotschedoff and Pachali, 2017). Finally, we summarize and discuss results in Section 6.

2 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. Gelfand, Smith, and Lee (1992) provide an overview of how to impose sign and order constraints based on truncated distributions using Gibbs sampling. Allenby, Arora, and Ginter (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. An alternative approach to including constraints is through the log-normal distribution (e.g., Allenby et al., 2014).

The critical role of constraints in a hierarchical model is 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 $\int \pi(a,\beta)\ p(\beta|\tau)\ d\beta = \mathbb{E}_{\beta|\tau}\left[\pi(a,\beta)\right] = \pi(a)$ 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 $Y = \begin{pmatrix} y_1 & \dots & y_i & \dots & y_N \end{pmatrix}$, 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:

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

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 even if *all* individual level price coefficients are *reliably* negative a posteriori.

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 biasing inferences about heterogeneity. Expected profits from action a are then estimated as:

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

However, as we illustrated in the introduction, 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.

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

The difference between this estimator and that defined in Equation 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 Equation 4 converges to that defined in Equation 3. When individual level posterior distributions come from diffuse individual level likelihoods, as usual, the estimator in Equation 4 will be very similar to that in Equation 2. Thus, parametric assumptions in the hierarchical prior distributions will be similarly influential.

If one wants to keep with the efficiency afforded by parametric or semi-parametric formulations of the hierarchical prior, it is therefore imperative to avoid the extrapolation of parametric assumptions into directions that violate prior theoretical knowledge using sign and order constraints. The goal is thus a hierarchical prior that both is maximally flexible regarding some aspects of the population distribution of preferences, and heavily constrained by theory regarding other aspects of this distribution.

Next we develop such a prior based on the log-normal distribution. The basic idea of using the log-normal distribution to implement sign and order constraints is not new (see e.g., Allenby et al., 2014). Our technical 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, 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.

Marginal-conditional decomposition

Our hierarchical prior starts with a standard normal distribution.[4] 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 $g : \mathbb{R}^k \to \mathbb{R}^k_c$ as the function that maps normally distributed variates β_i^* to sign and order constrained coefficients, β_i that enter the likelihood. We distinguish k_c "constrained" coefficients 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.

$$
\begin{aligned}
\beta_i^* &\sim N\big(\bar{\beta}^*, V_{\beta^*}\big), \text{ or} \\
\begin{pmatrix} \beta_i^{*c} \\ \beta_i^{*uc} \end{pmatrix} &\sim N\left(\begin{pmatrix} \mu_c^* \\ \mu_{uc}^* \end{pmatrix}, \begin{pmatrix} V_{\beta_{11}^*} & V_{\beta_{12}^*} \\ V_{\beta_{21}^*} & V_{\beta_{22}^*} \end{pmatrix}\right)
\end{aligned} \tag{5}
$$

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 Equation 5 in the form of a multivariate regression model that regresses unconstrained coefficients β_i^{*uc} on "constrained" coefficients β_i^{*c}:

$$
\begin{aligned}
B^{*uc} &= B^{*c}\,[z, \Gamma] + U \\
\beta^{*uc} &= (I_N \otimes [z, \Gamma]')\beta^{*c} + u
\end{aligned} \tag{6}
$$

Here, B^{*uc} and B^{*c} are matrices with k_{uc} and $k_c + 1$ columns, respectively, and N rows each, collecting unconstrained and "constrained" coefficients from individuals in the sample, where the first column in B^{*c} is a vector of ones. Γ is a $(k_c \times k_{uc})$ matrix of regression coefficients, z a vector of intercept coefficients of length k_{uc}, and $[\cdot]$ stacks objects *row*-wise such that $[z, \Gamma]$ denotes a $([k_c + 1] \times k_{uc})$ matrix with the intercept vector included in the first *row*. Finally, the second line in Equation 6 is the regression equation in vectorized form and $u := vec(U') \sim N(0, I_N \otimes \Sigma)$, where 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:

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

(7) $$B^{*c} = \iota(\mu_c^*)' + U_{V^*}$$

Here, ι is a $(N \times 1)$-vector of 1's and $u_{V^*} \sim N(0, I_N \otimes V^*)$ where V^* is the marginal variance-covariance matrix of constrained coefficients where $u_{V^*} := vec(U'_{V^*})$. The multivariate regression models in Equations 6 and 7 imply the following re-parameterization of the joint distribution of β_i^* from Equation 5:

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

The advantage of the re-parameterization in Equation 8 relative to the more standard parameterization in Equation 5 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 affecting the distribution of unconstrained parameters. At the same time, we can elect to be minimally informative about the distribution of unconstrained parameters governed by $[z, \Gamma]$ and Σ.

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 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).

	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

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

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^*} that in most applications would be considered as unduly informative as a prior for unconstrained coefficients where $\beta_i = \beta_i^*$. We further elaborate on this aspect in the simulation study in Section 3.

We use the fully conjugate prior for $([z, \Gamma], \Sigma)$ and the conditionally conjugate prior for (μ_c^*, V^*):

$$
\begin{aligned}
p([z,\Gamma],\Sigma) &= p([z,\Gamma]\,|\Sigma)p(\Sigma) \\
\gamma|\Sigma &\sim N(\bar{\gamma}, \Sigma \otimes A_{\Gamma}^{-1}),\ \gamma := vec([z,\Gamma]) \\
\Sigma &\sim IW(\nu_{\Sigma}, \bar{\Sigma}) \text{ and} \\
p(\mu_c^*, V^*) &= p(\mu_c^*)p(V^*), \\
\mu_c^* &\sim N(\bar{\mu}_c^*, A_{\mu_c^*}^{-1}) \\
V^* &\sim IW(\nu_{V^*}, \bar{V}^*)
\end{aligned} \tag{9}
$$

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 $\bar{\mu}_c^* = \begin{pmatrix} 0 & \dots & 0 \end{pmatrix}'$, $A_{\mu_c^*} = 0.1 I_{k_c}$, $\nu_{V^*} = k_c + 15$ as well as $\bar{V}^* = 0.5\nu_{V^*} I_{k_c}$, where I_{k_c} is the identity matrix of dimension $k_c \times k_c$ (see also 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.[5]

The fully conjugate prior for $([z,\Gamma],\Sigma)$ adjusts the influence of the subjective prior on $[z,\Gamma]$ 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, $\bar{\gamma},\ A_{\Gamma},\ \nu_{\Sigma},\ \bar{\Sigma}$.

Our marginal-conditional decomposition corresponds to directed acyclic graph in Figure 2 which shows that the hierarchical prior for "constrained coefficients", (μ_c^*, V^*), and that of unconstrained coefficients, $([z,\Gamma],\Sigma)$, are independent conditional on draws of "constrained" coefficients, B^{*c}. This conditional independence relationship gives rise to a Gibbs-sampler for (μ_c^*, V^*) conditional on "constrained" coefficients and subjective prior parameters, and another Gibbs-sampler for $([z,\Gamma],\Sigma)$ conditional on both "constrained" and unconstrained coefficients and subjective prior parameters (see Appendix A.1 for the explicit posterior distributions).

[5] In the latest release of the `R`-package `bayesm` (Version 3.1-0.1), Peter Rossi assesses subjective prior parameters for the joint distribution of "constrained" and unconstrained coefficients. While the resulting prior seems reasonable, `bayesm` relies on different default priors in the absence of constraints. The marginal-conditional decomposition we propose precisely resolves this conflict between prior preferences for "constrained" and unconstrained coefficients.

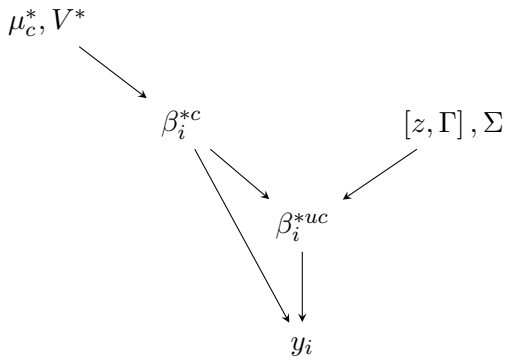

Figure 2: Marginal-conditional decomposition DAG.

Efficient MH-sampling

Next we discuss efficient sampling of individual level part worth coefficients $\{\beta_i^*\}$ based on pre-tuned proposal densities in a Metropolis-Hastings (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 delivers information about $\{\beta_i\}$ in closed form. However, our hierarchical prior is on the distribution of $\{\beta_i^*\}$; therefore, we need to account for this change-of-variables.

Following Rossi et al. (2005), we specify the proposal density of a random walk sampler as follows

$$\beta_i^{*cand} \sim N\left(\beta_i^*, c^2(H_i^* + (V_{\beta^*}^r)^{-1})^{-1}\right), \tag{10}$$

where $r \in \{1, \ldots, R\}$ is the r-th draw 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:

$$l_i^{\text{fract}}(\{y_i\}_{i=1}^N | g(\beta_i^*)) = MNL(y_i | g(\beta_i^*))^{1-w} \; MNL(\{y_i\}_{i=1}^N | g(\beta_i^*))^{w(T_i/\bar{T})} \tag{11}$$

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 Equation 12, 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.[6]

$$H_i^* \approx (J_g)' \, H_i J_g \tag{12}$$

[6] Appendix A.2 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 Equation 12.

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., the parameter value that maximizes the fractional likelihood in 11.

Appendix A.3 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, for two szenarios of individual observations (see Table 19). We note that these differences can magnify substantially in a hierarchical setting.

3 Simulation study

This section illustrates the benefits of our proposed marginal-conditional decomposition in the presence of sign and order constraints in a simulation exercise. Moreover, we elaborate on the drawbacks of using posterior means of individual level coefficients $\hat{\beta}_i = \int \beta_i p(\beta_i|Y, y_i) d\beta_i$ as a basis for market simulation and estimate the implied losses in profits when relying on this method for decision-making.

Setting

Suppose a monopolist wants to determine which one product to offer at what price to maximize profits. There are two attributes $A1$ and $A2$ at two possible levels $L1$ and $L2$ each, yielding four possible product configurations to choose from. 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 monopolist considers prices P in the range 0.50 to 3.00. The price coefficient is negative. We thus have the following set of constraints for every consumer $i = 1, \ldots, N$:

$$\begin{aligned} \beta_{+,i}, \beta_{++,i} &\geq 0 \\ \beta_{++,i} &\geq \beta_{+,i} \\ \beta_{p,i} &\leq 0 \end{aligned} \tag{13}$$

We generate heterogeneous consumer preferences obeying these constraints using the following transformation and distribution:

$$\beta^* = \begin{pmatrix} \beta^*_{+} \\ \beta^*_{++} \\ \beta^*_{p} \\ \beta^*_{uc_1} \\ \beta^*_{uc_2} \end{pmatrix} = g^{-1}(\beta) = \begin{pmatrix} \ln(\beta_{+}) \\ \ln(\beta_{++} - \beta_{+}) \\ \ln(-\beta_p) \\ \beta_{uc_1} \\ \beta_{uc_2} \end{pmatrix} \sim N\left(\bar{\beta}^*, V_{\beta^*}\right), \text{ with :}$$

$$\bar{\beta}^* = \begin{pmatrix} 0.5 & -0.5 & 0.8 & 2.5 & 2.5 \end{pmatrix}' \text{ and} \tag{14}$$

$$V_{\beta^*} = \begin{pmatrix} 0.4 & 0.1 & 0 & 0 & 0 \\ 0.1 & 0.2 & -0.15 & 0 & 0 \\ 0 & -0.15 & 0.4 & -0.05 & 0.05 \\ 0 & 0 & -0.05 & 2 & 0 \\ 0 & 0 & 0.05 & 0 & 4 \end{pmatrix}$$

	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 2: Summary of marginal distributions of data generating coefficients.

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 vary the amount of individual level information and investigate inference based on $T = 2,\ 4,\ 5,\ 10,\ 15$ and 100. Recall that many discrete choice studies in marketing barely reach one choice task per parameter to estimate at the individual level. The results in the range of $T = 2$ to $T = 5$ will therefore be most representative from a practical perspective. Intuitively, there will be a lot of uncertainty about individual parameters in that range and, consequently, the (hierarchical) prior will strongly influence the posterior location of individual coefficients. Results for $T = 10$ and $T = 15$ are meant to represent (unusually) informative individual level data, given the small dimensionality of our model. While the prior still matters in this range, the data start to become more and more informative about individual parameters such that the results from different approaches to generalizing to the population are expected to converge.

The results from $T = 100$ verify whether the model recovers data generating parameters and as a benchmark for results obtained with small T.

We remove the column pertaining to the first level of A_1 from the design matrix for identification.[7] 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 $\beta^{id}_{uc_2} = \beta_{uc_2} + \beta_{+}$.

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

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

Estimates of heterogeneity

We discussed before that the shrinkage of individual level posterior means in general reduces the number of sign and order violations and that this circumstance materially contributed to the popularity of market simulators based on posterior means. Table 4 illustrates the percentage of sign and order violations in an unconstrained model in our simulation example across the different forms of generalization and different T.[8] The percentage of sign violations is computed as the fraction of posterior means (posterior draws) that exhibit at least one order or sign violation.

	Posterior Means	Hierarchical Prior	Lower Level Model (n.s.)
T=2	1.5	20.5	20.2
T=4	2.1	17.2	16.6
T=5	2.5	15.8	13.9
T=10	0.2	8.5	7.0
T=15	0.2	6.7	5.1
T=100	0.1	6.4	2.4

Table 4: Percentage of order or sign violations in an unconstrained hierarchical Bayes model implied by different forms of generalization.

As apparent from Table 4, posterior means drastically reduce sign and order violations. The posterior of the hierarchical prior and lower level model (n.s.), on the other hand, both substantially support preferences that violate order or sign constraints. Thus, market simulations based on the unconstrained posterior of the hierarchical prior or the corresponding lower level model (n.s.) will be misleading and

[7] 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.

[8] The unconstrained model employs standard weakly informative subjective priors.

may produce results that lack face validity in practice. Market simulations based on posterior means could be more face valid but will be misleading because of a biased representation of heterogeneity in the market.

The results in Table 4 call for a constrained hierarchical prior distribution of heterogeneity in the population. Figure 3 illustrates the benefits of our proposed marginal-conditional decomposition of the hierarchical prior distribution (see Equations 6 to 8) compared to the standard formulation (see Equation 5) coupled with informative subjective prior settings as in e.g., Allenby et al. (2014) using the example of the unconstrained coefficient $\beta_{uc_2}^{id}$.[9]

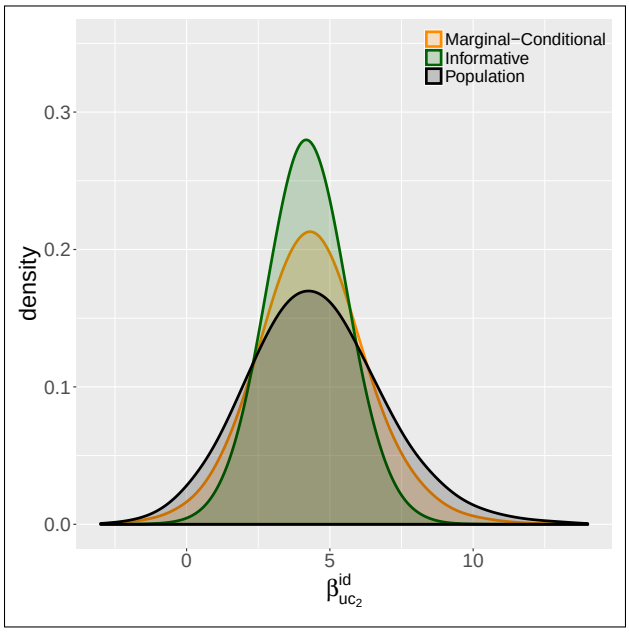

Figure 3: Posterior predictive population distributions of $\beta_{uc_2}^{id}$ using the marginal-conditional decomposition and the standard formulation ($T = 2$).

It is visually apparent that the model imposing informative priors on both constrained and unconstrained parameters underestimates the amount of preference heterogeneity in the unconstrained coefficient $\beta_{uc_1}^{id}$. The bias from unduly informative priors on unconstrained coefficients likely 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.[10] We illustrate the benefits of our marginal-conditional decomposition in the context of non-normal heterogeneity in Section 5.

Next, we further elaborate on the bias caused by using individual level posterior means as a representation of heterogeneity, but dropping the comparison between the marginal-conditional decomposition and the standard (constrained) hierarchical prior formulation to avoid clutter. Figure 4, qualitatively replicates the bias in the inferred population distributions of unconstrained coefficients illustrated in

[9]Note that the normal hierarchical prior for $\beta_{uc_1}^{id}$ and $\beta_{uc_2}^{id}$ used in estimation no longer exactly corresponds to the data generating heterogeneity distribution in this example. The data generating marginal distributions of $\beta_{uc_1}^{id}$ and $\beta_{uc_2}^{id}$ 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.

[10]Allenby et al. (2014) propose an even tighter IW-prior for the variance-covariance matrix in a mixture of normals model.

the introduction when relying on posterior means of random coefficients for $T = 4$ and $T = 15$. The bias is still substantial, even with $T = 15$ and only four parameters at the individual level. Also note the close correspondence between the posterior predictive densities from the hierarchical prior and lower level (n.s.).

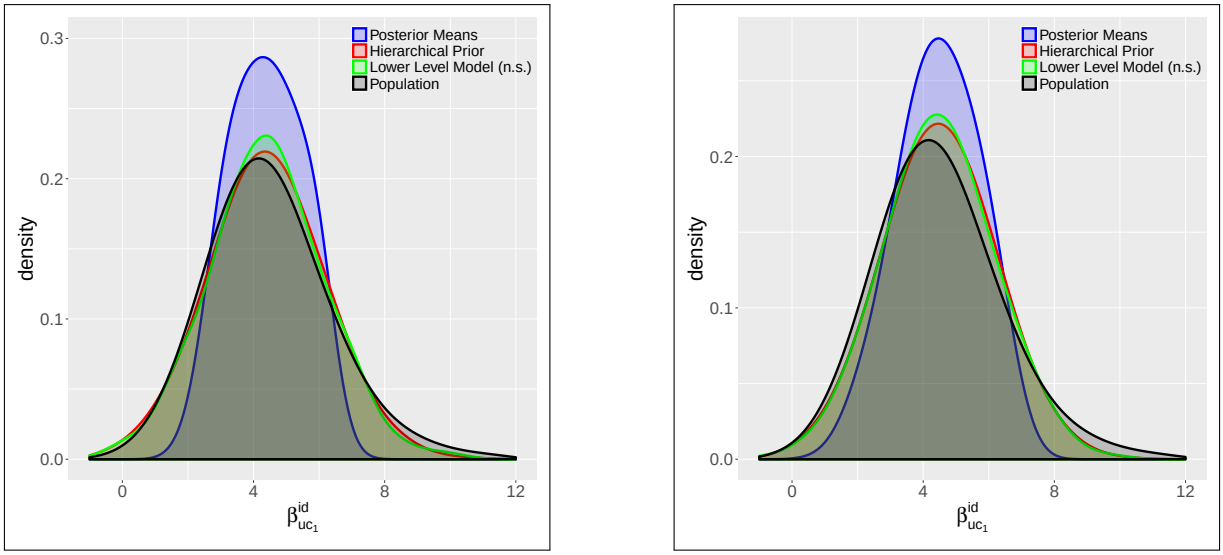

Figure 4: Posterior predictive population distributions using posterior means, the posterior of the hierarchical prior and lower level model (n.s.) for $\beta^{id}_{uc_1}$ and $T = 4$ as well as $T = 15$.

A new dimension to the bias from relying on the collection of individual level posterior means is illustrated in Figure 5, in the context of constrained coefficients. Figure 5 displays the true and the inferred distributions of the sign-constrained price coefficient in the population based on $T = 4$ and $T = 15$. While the distribution implied by the collection of individual level posterior means is again more concentrated than the data generating distribution and that implied by the posterior of the hierarchical prior, it is also displaced to the left in the direction of a higher overall price sensitivity.

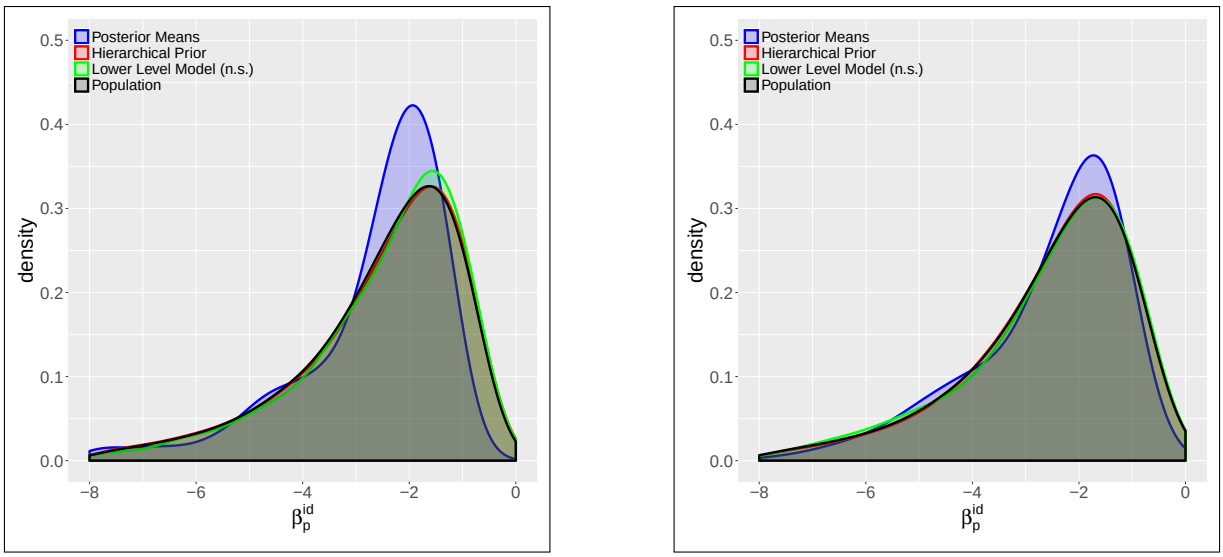

Figure 5: Posterior predictive population distributions using posterior means, the posterior of the hierarchical prior and lower level model (n.s.) for the price coefficient and $T = 4$ as well as $T = 15$.

The combination of the increased concentration and the displacement towards higher price sensitivity leads to underestimating the percentage of less price sensitive consumers with implications for optimal pricing. The distributions implied by the posterior of the hierarchical prior and lower level model (n.s.), on the other hand, do not exhibit this bias and better recover the data generating population distribution both when $T = 4$ and when $T = 15$.

The source of the mean displacement of the distribution inferred from the collection of individual level posterior means is the skewness in the individual level posterior distributions of the price coefficient, and specifically the relationship between individual level posterior uncertainty and the skewness in the individual level posterior. Figure 6 plots posterior means against the posterior variance in individual level posterior distributions of the price coefficient for the case of $T = 4$. The figure shows that posterior means translate increased individual level posterior uncertainty (on the x-axis) into increased inferred price sensitivity (on the y-axis).

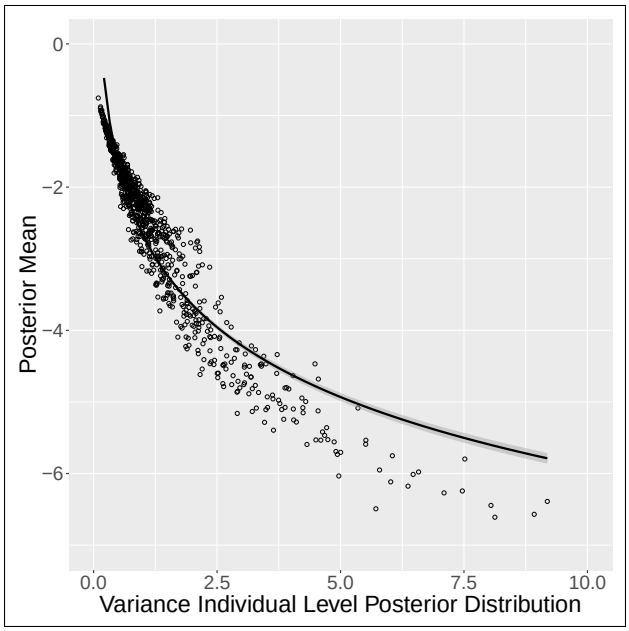

Figure 6: Relation between individual level posterior means (y-coordinate) and variance of individual level posterior distributions (x-coordinate) illustrated for β_p^{id} and $T = 4$.

Fitting the model $\hat{\beta}_{p,i}^{id} = a + b \ln\Big(\text{Var}(\beta_{p,i}^{id}|Y, y_i)\Big)$ to the data in Figure 6, we obtain that doubling the individual level posterior variance approximately decreases the price coefficient by -1, on average in the case of $T = 4$. We note that means and variances are correlated for all constrained distributions whenever the constraint matters.

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 three 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 (HP) and hit rates (HR), holdout log-likelihoods are the more 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, the collection of individual level posterior means, and lower level model (n.s.), HL of individual h's choices are defined as in 15, 16, and 17, respectively. In each case $HLL(Y_{\text{hold}}) = \sum_{h=1}^{H} ln(HL(y_h))$.

$$HL(y_h) = \int MNL(y_h|g(\beta_h^*))p(\beta_h^*|\bar{\beta}^*, V_{\beta^*})p(\bar{\beta}^*, V_{\beta^*}|Y_{\text{train}})\ d(\beta_h^*, (\bar{\beta}^*, V_{\beta^*})) \tag{15}$$

$$HL(y_h) = \frac{1}{N}\sum_{i=1}^{N} MNL(y_h|\hat{\beta}_i), \tag{16}$$

$$HL(y_h) = \frac{1}{N}\sum_{i=1}^{N} \int MNL(y_h|g(\beta_h^*))p(\beta_h^*|y_i, \bar{\beta}^*, V_{\beta^*})p(\bar{\beta}^*, V_{\beta^*}|Y_{\text{train}})\ d(\beta_h^*, (\bar{\beta}^*, V_{\beta^*})) \tag{17}$$

Table 5 illustrates the resulting holdout log-likelihood estimates implied by the three forms of generalization. The numbers are directly comparable as they are computed using a fixed data set comprising $T = 15$ choices made by each of $H = 500$ holdout respondents from the same population. The choice design is the same as that used to generate the calibration data.

	Posterior Means	Hierarchical Prior	Lower Level Model (n.s.)
T=2	-7642	-7510	-7511
T=4	-7546	-7485	-7485
T=5	-7533	-7482	-7482
T=10	-7508	-7475	-7476
T=15	-7500	-7472	-7480
T=100	-7481	-7470	-7477

Table 5: Predictive performance (holdout log-likelihoods) of different forms of generalization for different T. ***MAX: Should the add predictions obtained with the Rossi prior?***

Generalizations to the population—as operationalized by predictions of choices by our 500 holdout respondents—based on the hierarchical prior and lower level model (n.s.) outperform those using the collection of individual level posterior means for all T. In particular, we find large predictive performance differences (between 50 and more than 100 log-likelihood points) for the practically relevant cases in the range of $T = 2$ to $T = 5$.[11] As expected, this difference becomes smaller as the individual level data become more informative about individual level preferences. However, we still see sizeable predictive performance differences between posterior means on the one side and the hierarchical prior or lower level model (n.s.) on the other side for the cases of $T = 10$ and $T = 15$.

To gauge the managerial relevance of these differences, we return to the monopolist's decision problem of which product to offer at what price, as a function of expected profits. Our monopolist seeks to find

[11] We do not imply that $T = 2$ to $T = 5$ are per se practically relevant but that ratios of 4 individual level parameters to 2, 4, and 5 individual-level observations are.

the optimum out of the set of possible actions denoted as $A = \{(A1_{L1}, A2_{L1}, P), \ldots, (A1_{L1}, A2_{L2}, P), \ldots, (A1_{L2}, A2_{L2}, P)\}$. We define the optimal decision under the true data generating heterogeneity distribution as a_{opt} that solves 1, where $\text{MS}(a, \beta) = \Pr(a|\beta) = \frac{\exp(x_a'\beta)}{1+\exp(x_a'\beta)}$ is the logit choice probability given action a from the monopolist's perspective. Similarly, we denote a_{hp}, a_{pm}, and a_{llmns} as the optimal actions derived from the hierarchical prior, the collection of individual level posterior means, and lower level model (n.s.), respectively.

The optimal decision also depends on the associated variable costs $C(a)$. To make sure that our comparison does not hinge on the specifics of a particular variable cost setting, we systematically vary costs. Without loss of generality we assume that only the first attribute has non-zero production costs and that producing the second, more preferred level is three times as costly as producing the first. Finally, we create a cost grid with 2000 scenarios and thus induce variation in the choice of optimal products independent of the analyst's assessment of the preferences in this market.

	Level 1	Level 2
Minimum	0.0	0.0
Mean	0.3	1.0
Maximum	0.7	2.0

Table 6: Minimum, mean and maximum costs of levels of attribute one.

Table 6 summarizes minimum, mean and maximum of the grid specified for $c_{A1_{L1}}$ and $c_{A1_{L2}}$. Cost differences vary over this grid. At the beginning, both levels almost have the same low costs (approximately equal to zero at the minimum). In this situation, the choice of the second, more preferred level of attribute 1 (see Table 2) for profit maximization is straightforward. Once costs start to increase, it becomes more and more costly for the monopolist to offer the second, more preferred level of this attribute. In combination with the dependencies between price sensitivity and preferences for the levels of the second attribute, this collection of cost scenarios supports the full range of possible products as optimal, depending on the specific cost setting.

We measure performance by computing the loss in expected profits relative to $a_{\text{opt,k}}$ computed at the true, data generating distribution of preferences for each cost scenario $k = 1, \ldots, K$ (see Equation 1). Equation 18 shows the corresponding expression for the optimal action obtained based on the collection of individual level posterior means at the k-th cost scenario, where $\pi()$ is computed with respect to the true heterogeneity distribution.

$$L(a_{\text{pm},k}) = \frac{\pi(a_{\text{opt},k}) - \pi(a_{\text{pm},k})}{\pi(a_{\text{pm},k})} \tag{18}$$

Table 7 summarizes the distribution of relative losses from the different approaches to generalizing to the population conditional on how many choice observations per respondent (T) are in the calibration data. Actions computed using the posterior of the hierarchical prior or lower level model (n.s.) both completely dominate actions inferred based on the collection of individual level posterior means up to $T = 4$. The *maximum* relative loss incurred by the former methods is *smaller* than the *minimum* relative loss incurred when relying on the collection of individual level posterior means; across cost

scenarios, average relative losses from using this latter method are between 9.67% and 1.81% for the practically most relevant cases in the range of $T = 2$ and $T = 5$ respectively.

	Posterior Means			Hierarchical Prior			Lower Level Model (n.s.)		
	Minimum	Mean	Maximum	Minimum	Mean	Maximum	Minimum	Mean	Maximum
T=2	4.643	9.672	13.839	0.629	0.956	1.353	0.574	1.286	1.625
T=4	1.036	2.856	4.560	0.015	0.037	0.065	0.005	0.065	0.117
T=5	0.516	1.809	3.066	0.000	0.001	0.184	0.000	0.008	0.549
T=10	0.119	0.787	1.534	0.003	0.029	1.005	0.025	0.122	1.317
T=15	0.027	0.349	1.128	0.010	0.033	0.876	0.000	0.009	0.330
T=100	0.000	0.017	0.315	0.002	0.014	0.318	0.000	0.014	0.311

Table 7: Distribution of relative losses in expected profits (in %) using posterior means, the posterior of the hierarchical prior, and lower level model (n.s.) for different T.

Optimal actions based on the collection of individual level posterior means are still markedly inferior when T is 10 or even 15, with average relative losses of 0.79% and 0.35%. Average relative losses from relying on (the posterior of) the hierarchical prior amount to only 0.03% in this case, and those from using lower level model (n.s.) to about 0.12% and 0.01% with $T = 10$ and $T = 15$, respectively. Again, the differences between the three approaches to generalizing to a population only vanish once T becomes unrealistically large, allowing for precise individual level inferences. Based on this illustration, we conclude that the differences between approaches to generalizing to the population from a hierarchical model—and specifically the disadvantages of relying on the collection of individual level posterior means—are material for managerial decisions. However, in order to avoid the pitfalls from using individual level posterior means as a representation of the population preference distribution, constrained hierarchical prior distributions are required. And, in order to properly accommodate prior distributional differences between constrained and unconstrained coefficients in the hierarchical prior, we need the proposed marginal-conditional decomposition of the hierarchical prior. Next we illustrate the empirical relevance of our arguments in two case studies.

4 Tablet PC preferences

Our first empirical application uses data from a commercial discrete-choice conjoint study investigating demand for tablet PCs ("tablets"). Table 8 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.

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

Table 8: Attributes and levels in the tablet experiment.

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.

The original goal of the 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.

Many of the attributes and levels in Table 8 are such that one can expect every respondent to strictly prefer one level over another level, everything else equal. Table 9 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 seems unreasonable 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, 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 2 to specify suitable subjective priors.

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	$\beta_{\text{CB}_{150}} \geq \beta_{\text{CB}_{100}} \geq \beta_{\text{CB}_{50}} \geq \beta_{\text{CB}^-}$

Table 9: Restricted attributes and constraints imposed on levels.

We run the MCMC sampler using the tuned random walk proposal from Section 2 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. ***MAX: Do we really have to burn of the majority of draws here?*** 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.[12]

Figure 7 visually compares the marginal posterior population densities of coefficients measuring preferences for levels of the cash back attribute. Cash back is the amount of money a customer receives after purchase upon submitting the sales receipt to the manufacturer. 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.[13]

The upper left panel of Figure 7 shows inferred population preference distributions for 50€ cash back relative to no cash back. The bias incurred when constructing this distribution from posterior means of individual level coefficients is clearly visible. 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 distributions implied by the posterior of the hierarchical prior or by lower level model (n.s.), which are visually very similar, 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.

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

[13]We illustrate and summarize the respective marginal posterior densities from an unconstrained model estimated using standard weakly informative subjective priors (e.g., Rossi et al., 2005) in Figure 10 and Table 20 in Appendix A.4. The results from the unconstrained model clearly violate basic intuition regarding the e.g., the relative attractiveness of cash back levels.

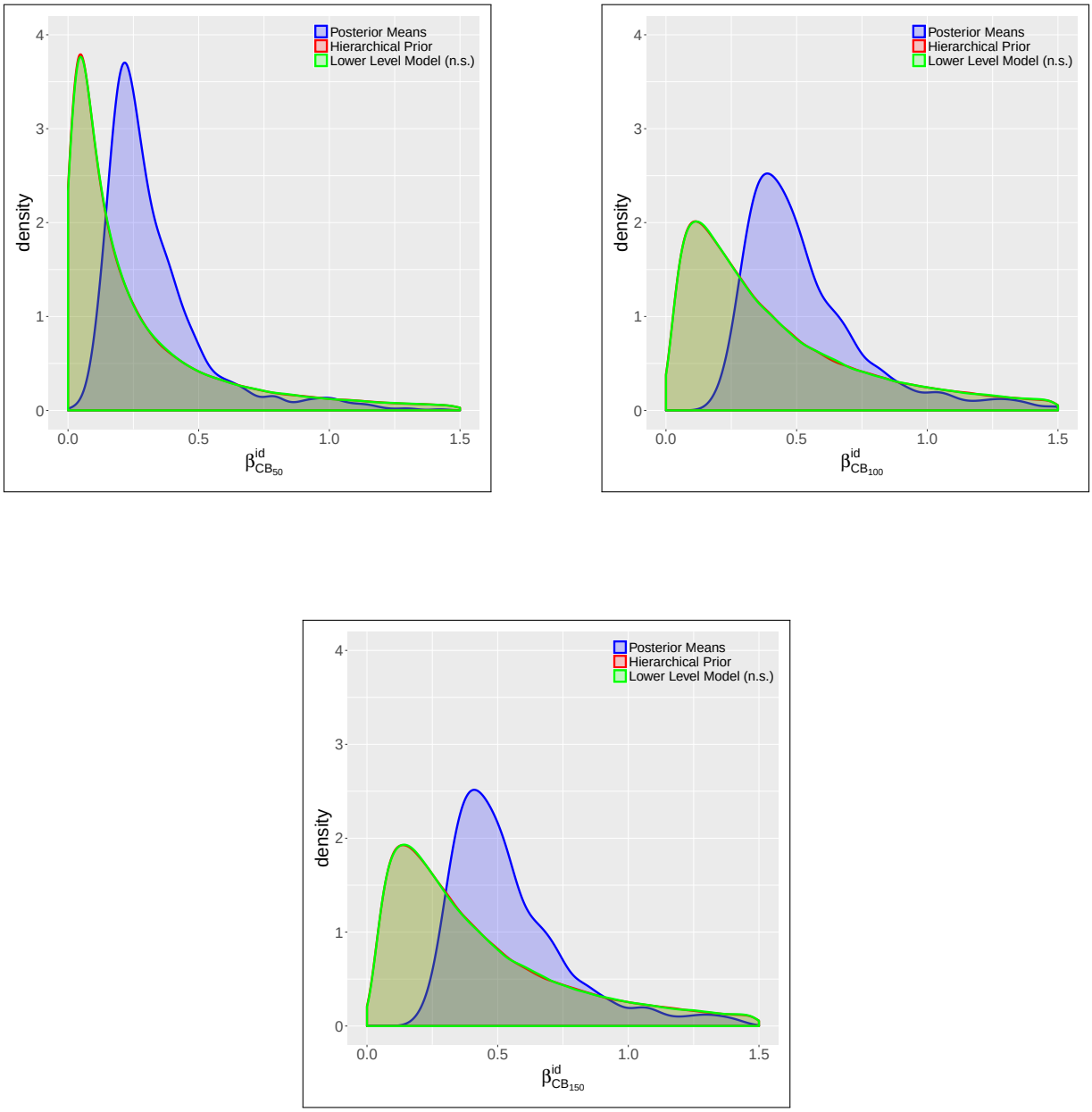

Figure 7: Posterior predictive population densities for the levels of the cash back attribute using posterior means, the posterior of the hierarchical prior, and lower level model (n.s.).

Figure 8 illustrates inferred population preference distributions for brand A and C. We see—in line with the simulation results reported earlier—that the collection of individual level posterior means underestimates the degree of taste heterogeneity for these two brands. The population distributions inferred from the posterior of the hierarchical prior and lower level model (n.s.) are very similar.

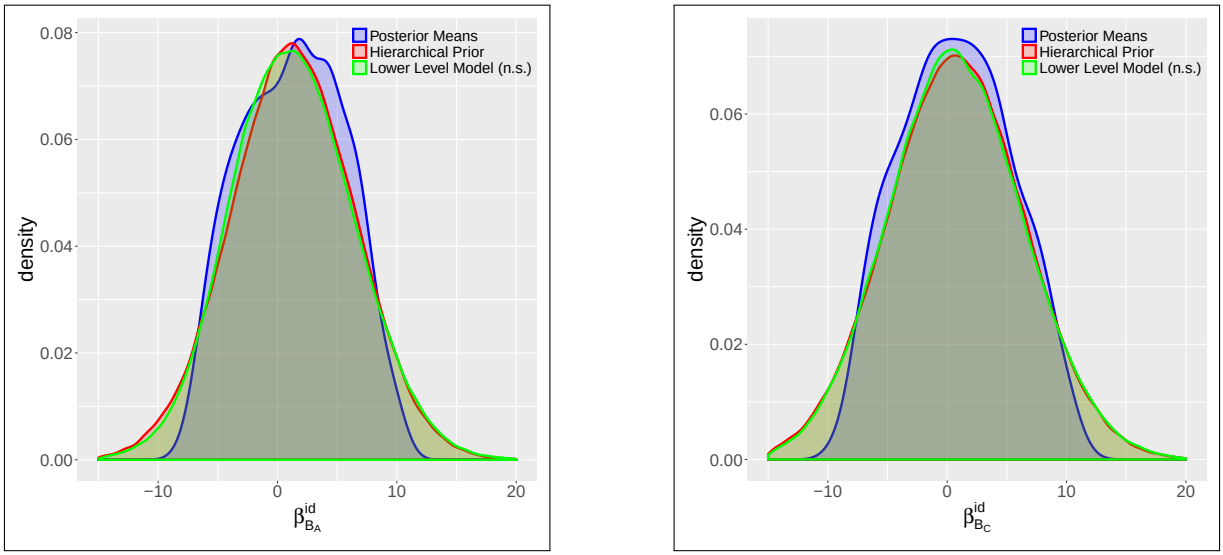

Figure 8: Posterior predictive population densities of brand A (left panel) and C (right panel) coefficients using posterior means, the posterior of the hierarchical prior, and lower level model (n.s.).

Next we evaluate the predictive performance of the population preference distributions inferred from the collection of individual level posterior means, the posterior of the hierarchical prior, and lower level model (n.s.) 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 (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):

$$\begin{aligned} \text{CV}_{\text{HLL}}(Y) &= \frac{1}{K}\sum_{k=1}^{K}\sum_{y_h \in Y^k_{\text{hold}}} \text{HLL}(A(Y^k_{\text{train}}), y_h) \\ &= \frac{1}{K}\sum_{k=1}^{K} \text{HLL}(A(Y^k_{\text{train}}), Y^k_{\text{hold}}), \end{aligned} \tag{19}$$

$\text{HLL}(A(Y^k_{\text{train}}), y_h)$ 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 Equations 15 - 17). 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, the posterior of the hierarchical prior, or based on lower level model (n.s.).

	Posterior Means	Hierarchical Prior	Lower Level Model (n.s.)
Fold 1	-2499	-2495	-2466
Fold 2	-2606	-2630	-2598
Fold 3	-2810	-2784	-2755
Fold 4	-2761	-2744	-2718
Fold 5	-3512	-3509	-3454
Mean	-2837	-2832	-2798

Table 10: Predictive performance (holdout log-likelihoods, five-fold cross-validation) of different forms of generalization.

Table 10 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, the posterior of the hierarchical prior, and lower level model (n.s.), the latter outperforms the former two not only on average but also in every single fold. Predictions based on the posterior of the hierarchical prior are better than those using the collection of individual level posterior means in four out of the five folds and are better on average, however only by a small margin.[14]

Based on these results, market predictions that rely on the collection of individual level posterior means or on the posterior of the hierarchical prior only appear suboptimal. While the inferior performance of the collection of individual level posterior means follows from our earlier discussion, the horse-race between the posterior of the hierarchical prior and lower level model (n.s.) is an empirical issue in applications. Both methods critically depend on a reasonable specification of the hierarchical prior distribution in the usual large N, small T setting. However, lower level model (n.s.), by construction, can depart from the predictions implied by the posterior of the hierarchical prior to the extent that individual choice vectors $\{y_i\}$ in the calibration data are informative about mis-specifications in the functional form of the hierarchical prior.

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 and configured as follows:

	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

Table 11: Specification of products offered by brand A's competitors.

[14]The five-fold cross-validation log-likelihoods using the unconstrained model are -2997, -2894, and -2867 based on posterior means, the posterior of the hierarchical prior, and lower level (n.s.), respectively. Constraining the hierarchical prior therefore improves the predictive performance of the model, and regardless of how the model is translated into posterior predictions.

To more generally capture differences between optimal actions implied by the three 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.

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

Table 12: Minimum, mean and maximum of product-specific costs illustrated for five cost scenarios.

Table 12 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 13 summarizes the distribution of brand A's expected percentage losses incurred by relying on the collection of individual level posterior means and the posterior of the hierarchical prior relative to lower level model (n.s.) across cost scenarios. All losses are computed conditional on the market representation by lower level model (n.s.) for the optimal actions a_{pm}, a_{hp}, and a_{llmns}. (Recall that lower level model (n.s.) produced the best market level predictions.)

	Minimum	Mean	Maximum
Posterior Means	1.162	6.683	12.193
Hierarchical Prior	0.011	0.527	0.772

Table 13: Percentage losses from using posterior means, the posterior of the hierarchical prior across cost scenarios relative to optimal actions from lower level model (n.s.).

We see that optimization results that rely on the collection of individual level posterior means to represent market preferences are clearly inferior. The maximum percentage loss when relying on the posterior of the hierarchical prior is smaller than the minimum percentage loss from the collection of individual level posterior means, and the average percentage loss of 6.68% from using this latter method seems substantial.

5 Preferences for fresh hen's eggs

Our second empirical application relies on Nielsen data on fresh hen's eggs purchases of German households used in Kotschedoff and Pachali (2017) and serves to illustrate the practical relevance of the proposed marginal-conditional decomposition of the hierarchical prior. For clarity of exposition and to conserve space, we do not again illustrate the drawbacks associated with relying on individual posterior means in this example. In Germany eggs are differentiated in terms of animal welfare as illustrated in Table 14.

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.

Table 14: Main differences between egg breeding categories.

There is an EU-wide requirement to state the breeding category on egg packages and to additionally print a code on each single egg indicating origin and breeding category since 2004. And consumers associate the four breeding categories with different quality levels: battery eggs $\precsim$ barn eggs $\precsim$ free-range eggs, and $\precsim$ organic eggs. In 1999 the EU decided that all member states had to ban the production of battery eggs by 2012 and Germany already implemented the ban in 2010. Kotschedoff and Pachali (2017) 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.[15]

The demand model they employ 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

$$U_{iglt} = \gamma_{i,g} + \alpha_i p_{glt} + \beta_i \mathbf{1}\{units_g = 6\} + \psi_{i,l} + \varepsilon_{iglt}, \tag{20}$$

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, $u_{iglt} = 0$. As standard in the literature, ε_{iglt} follows a type I extreme value distribution.

Kotschedoff and Pachali (2017) argue that a flexible estimation of the retail chain preference coefficients $\{\psi_{i,l}\}$ is particularly important in their demand specification to alleviate or even prevent the bias caused by the full information assumption implicit to Equation 20: 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

[15] 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.

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. Finally, they use the inferred information about $\{\psi_{i,l}\}$ to account for competition among retail chains in a supply side model.

Here, we rely on the simplified demand framework in Equation 20 to illustrate the benefits of our marginal-conditional decomposition model as developed in Section 2.[16] This application again represents the prototypical situation with a mix of constrained and unconstrained coefficients in a hierarchical model. While we cannot constrain preferences for the retail chains and the battery egg taste coefficient, which measures preferences for battery eggs over the outside good, a priori, 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 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 15. Preferences for the four different egg labels should satisfy the quality ordering implied by Table 14 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.

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

Table 15: Restricted attributes and constraints imposed on levels.

Table 16 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 that motivates the use of a hierarchical model is due to the relatively small amount of within quality tier price variation as compared to price variation across quality tiers.

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

Table 16: Distribution of the number of egg purchase incidents across $N = 498$ households used in the estimation sample.

[16] We do not discuss more complex model specifications, such as models controlling for seasonality or regime changes, that might be useful to rule out endogeneity concerns as they are not relevant for the problem we are illustrating here.

We compare our model (see Equations 6 to 8) to the standard formulation (see Equation 5) coupled with informative subjective prior settings as in e.g., 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. We draw a random subsample of $N = 498$ households and estimate a model with a five components mixture of normals prior (as in Kotschedoff and Pachali (2017)) under these two different subjective prior settings.[17]

Figure 9 shows posterior predictive population distributions for the (unconstrained) battery egg coefficient as well as the coefficient measuring preferences for retail chain 5.[18] Both graphs in Figure 9 confirm the finding from the simulation study in Section 3: By imposing an informative prior on all coefficients (that is really needed for the constrained coefficients only) the standard formulation underestimates preference heterogeneity for the unconstrained coefficients. This is particularly apparent in the right panel where the marginal posterior from the standard hierarchical prior with informative subjective prior choices fails to accommodate extremely negative preferences for retail chain 5 in the left tail.

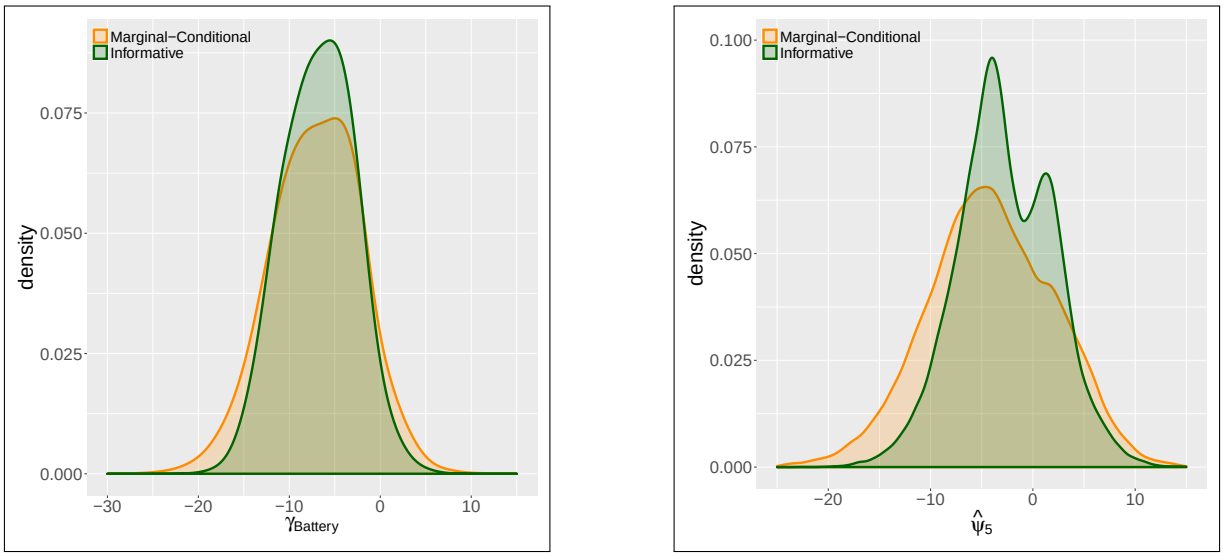

Figure 9: Posterior predictive population distributions using the marginal-conditional decomposition model and a standard model with informative priors for the battery egg coefficient (left panel) as well as the preference coefficient of the fifth retail chain (right panel).

Table 17 summarizes variances of marginal posterior predictive densities of unconstrained coefficients and verifies that the differences across the two subjective prior specifications are substantial.

[17] 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.

[18] We estimate individual retail chain preferences relative to a baseline chain in order identify the likelihood, 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.

	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 17: Variances of the marginal posterior densities of the estimated unconstrained preference coefficients implied by the marginal-conditional decomposition model and a standard model with informative priors.

Finally Table 18 compares model fit based on the Newton-Raftery estimator of the log marginal likelihood. As one may expect, the indistinctively informative specification of the standard prior translates into inferior fit.

Marginal-Conditional	Informative
-38853	-39001

Table 18: Comparison of log marginal likelihood values across model specifications.

6 Discussion

Hierarchical models have become ubiquitous in marketing because of the common data format that contains a small number of individual level observations (T), each from a relatively much larger number of observational units (N). Bayesian methods facilitate inference based on these models. As evidenced by current applications in industry and applied academic research, analysts value the efficiency afforded by fully or semi-parametric hierarchical prior formulations but remain skeptical about the value of these priors as faithful and economically meaningful representations of preference distributions in populations. As a consequence, the prevailing practice in market simulation relies on the collection of individual level posterior means as a representation of population preferences.

We demonstrate that this practice results in biased inference about optimal actions whenever individual level likelihoods are diffuse, precisely the situation that motivates the use of hierarchical models. We identify sign- and order constraints derived from prior economic arguments as a requirement for more faithful and economically meaningful representations of preference distributions, and rely on the log-normal distribution to impose these constraints. In this context, we develop a marginal-conditional decomposition of the hierarchical prior distribution that greatly facilitates the formulation of sensible subjective priors for the hierarchical prior distribution. Specifically, the proposed marginal-conditional decomposition resolves the conflict between wanting to be much more informative about the hierarchical prior distribution of constrained coefficients relative to that of unconstrained coefficients. It 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. We then 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 demonstrate that population distributions constructed from the collection of individual level posterior *distributions* (not means), referred to as lower level model (n.s.) in the paper, are maybe surprisingly close to those inferred from the posterior of the hierarchical prior in practically relevant settings. Therefore, meaningful inferences based on the collection of individual level posterior distributions too critically depend on a well specified hierarchical prior distribution. The advantage of inference based on the collection of individual level posterior distributions as compared to inference based on the posterior of the hierarchical prior is that the former can depart from the parametric assumptions in the hierarchical prior. A practical disadvantage is the formidably high dimensionality of the collection of individual level posteriors when N is large and the individual level model complex. The posterior of the hierarchical prior is much easier to store and share, and simulating, i.e., 'forecasting' individual preferences from the posterior of the hierarchcial prior is essentially costless.

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 populations 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 approach based on the multivariate log-normal distribution developed in this paper. However, a truncated normal hierarchical prior poses formidable computational problems already in the simple case of a single sign constraint (see Boatwright et al., 1999) and there does not appear to be a substantive reason to prefer one approach over the other, after accommodating differentially informative subjective priors for constrained coefficients in the log-normal case.

References

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

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

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

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

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

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

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

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Boatwright, P., R. McCulloch, and P. E. Rossi (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., G. J. Hitsch, and P. E. Rossi (2009). Do switching costs make markets less competitive? *Journal of Marketing Research 46*(4), 435–445.

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

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

Elrod, T. (2001). Recommendations for validation of choice models. In *Proceedings of the Sawtooth Software Conference*, pp. 225–243.

Evgeniou, T., M. Pontil, and O. Toubia (2007). A convex optimization approach to modeling consumer heterogeneity in conjoint estimation. *Marketing Science 26*(6), 805–818.

Ferjani, M., K. Jedidi, and S. Jagpal (2009). A conjoint approach for consumer- and firm-level brand valuation. *Journal of Marketing Research 46*(6), 846–862.

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

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

Huber, J., B. K. Orme, and R. Miller (1999). Dealing with product similarity in conjoint simulations. Research paper series, Sawtooth Software.

Jedidi, K., S. Jagpal, and P. Manchanda (2003). Measuring heterogeneous reservation prices for product bundles. *Marketing Science 22*(1), 107–130.

Kotschedoff, M. J. W. and M. J. Pachali (2017). Higher minimum quality standards and redistributive effects on consumer welfare.

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

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

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

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

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

Rossi, P. E. (2014). *Bayesian Non- and Semi-parametric Methods and Applications*. Princeton University Press.

Rossi, P. E., G. M. Allenby, and R. McCulloch (2005). *Bayesian statistics and marketing*. John Wiley & Sons.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (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.

Sonnier, G., A. Ainslie, and T. Otter (2007). Heterogeneity distributions of willingness-to-pay in choice models. *Quantitative Marketing and Economics 5*(3), 313–331.

A Appendix

A.1 Posterior distributions: log-normal prior

The posteriors associated with the priors in Equation 9 are (see e.g., Rossi et al., 2005):

$$
\begin{aligned}
&\Sigma|B^{*uc}, B^{*c} \sim IW\left(\nu_\Sigma + N, \bar{\Sigma} + S_\Sigma\right) \\
&\gamma|B^{*uc}, B^{*c}, \Sigma \sim N\left(\tilde{\gamma}, \Sigma \otimes \left((B^{*c})'B^{*c} + A_\Gamma\right)^{-1}\right), \text{ with} \\
&\tilde{\gamma} := vec(\tilde{\Gamma}),\ \tilde{\Gamma} = \left((B^{*c})'B^{*c} + A_\Gamma\right)^{-1}\left((B^{*c})'B^{*c}\hat{\Gamma} + A_\Gamma\bar{\Gamma}\right), \\
&\hat{\Gamma} = \left((B^{*c})'B^{*c}\right)^{-1}(B^{*c})'B^{*uc}, \text{ and} \\
&S_\Sigma = \left(B^{*uc} - B^{*c}\tilde{\Gamma}\right)'\left(B^{*uc} - B^{*c}\tilde{\Gamma}\right) + \left(\tilde{\Gamma} - \bar{\Gamma}\right)'A_\Gamma\left(\tilde{\Gamma} - \bar{\Gamma}\right)
\end{aligned} \tag{21}
$$

$$
\begin{aligned}
&\mu_c^*|B^{*c}, V^* \sim N\left(\tilde{\mu}_c^*, \tilde{A}_{\mu_c^*}\right) \\
&V^*|B^{*c}, \mu_c^* \sim IW\left(\tilde{\nu}_{V^*}, \tilde{V}^*\right) \text{ with} \\
&\tilde{A}_{\mu_c^*} = \left(N(V^*)^{-1} + A_{\mu_c^*}\right)^{-1}, \\
&\tilde{\mu}_c^* = \tilde{A}_{\mu_c^*}\left(\left(\iota' \otimes (V^*)^{-1}\right)\beta^{*c} + A_{\mu_c^*}\bar{\mu}_c^*\right), \\
&\tilde{\nu}_{V^*} = \nu_{V^*} + N \text{ and } \tilde{V}^* = \bar{V}^* + \left(B^{*c} - \iota(\mu_c^*)'\right)'\left(B^{*c} - \iota(\mu_c^*)'\right)
\end{aligned} \tag{22}
$$

A.2 Exact Hessian of transformed variates

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

$$
H_i^* = \frac{\partial^2 l_i}{\partial\beta_i^*\partial\beta_i^{*\prime}}, \tag{23}
$$

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

Taking first derivative yields

$$
\frac{\partial l_i}{\partial\beta_i^{*\prime}} = \frac{\partial l_i}{\partial g(\beta_i^*)'}\frac{\partial g(\beta_i^*)}{\partial\beta_i^{*\prime}}, \tag{24}
$$

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 24 can be expressed as

$$\left[\frac{\partial l_i}{\partial \beta_i^{*\prime}}\right]_j = \left[\bar{l}_i\right]_j := \nabla l_i J_g^j, \tag{25}$$

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

$$H_i^* = \begin{pmatrix} \frac{\partial \left[\bar{l}_i\right]_1}{\partial \beta_i^*} & \cdots & \frac{\partial \left[\bar{l}_i\right]_j}{\partial \beta_i^*} & \cdots & \frac{\partial \left[\bar{l}_i\right]_k}{\partial \beta_i^*} \end{pmatrix}, \tag{26}$$

with:

$$\frac{\partial \left[\bar{l}_i\right]_j}{\partial \beta_i^*} = J_g' H_i J_g^j + \frac{\partial J_g^j}{\partial \beta_i^*} \nabla l_i' \tag{27}$$

A.3 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:

$$\begin{aligned} \beta^* = g^{-1}(\beta) &= \begin{pmatrix} \beta_1 & \ln(\beta_2 - \beta_1) & \ln(\beta_3 - \beta_2) & \ln(\beta_4 - \beta_3) & \ln(\beta_5 - \beta_4) \end{pmatrix}' \\ &= \begin{pmatrix} -1 & 0.2 & 0.5 & -0.1 & -0.5 \end{pmatrix}' \end{aligned} \tag{28}$$

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 $\begin{pmatrix} 0.01 & 0.06 & 0.11 & \ldots & 1.46 \end{pmatrix}$. This parameter is set to its default value of $c^2 = 1$ (see Rossi et al., 2005) in our proposed tuning scheme.

	$T = 20$		$T = 1000$	
	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

Table 19: Numerical efficiency of MCMC, standard versus proposed tuning, $N = 1$, $T = 20$ and $T = 1000$.

A.4 Tablet PC preferences in an unconstrained model

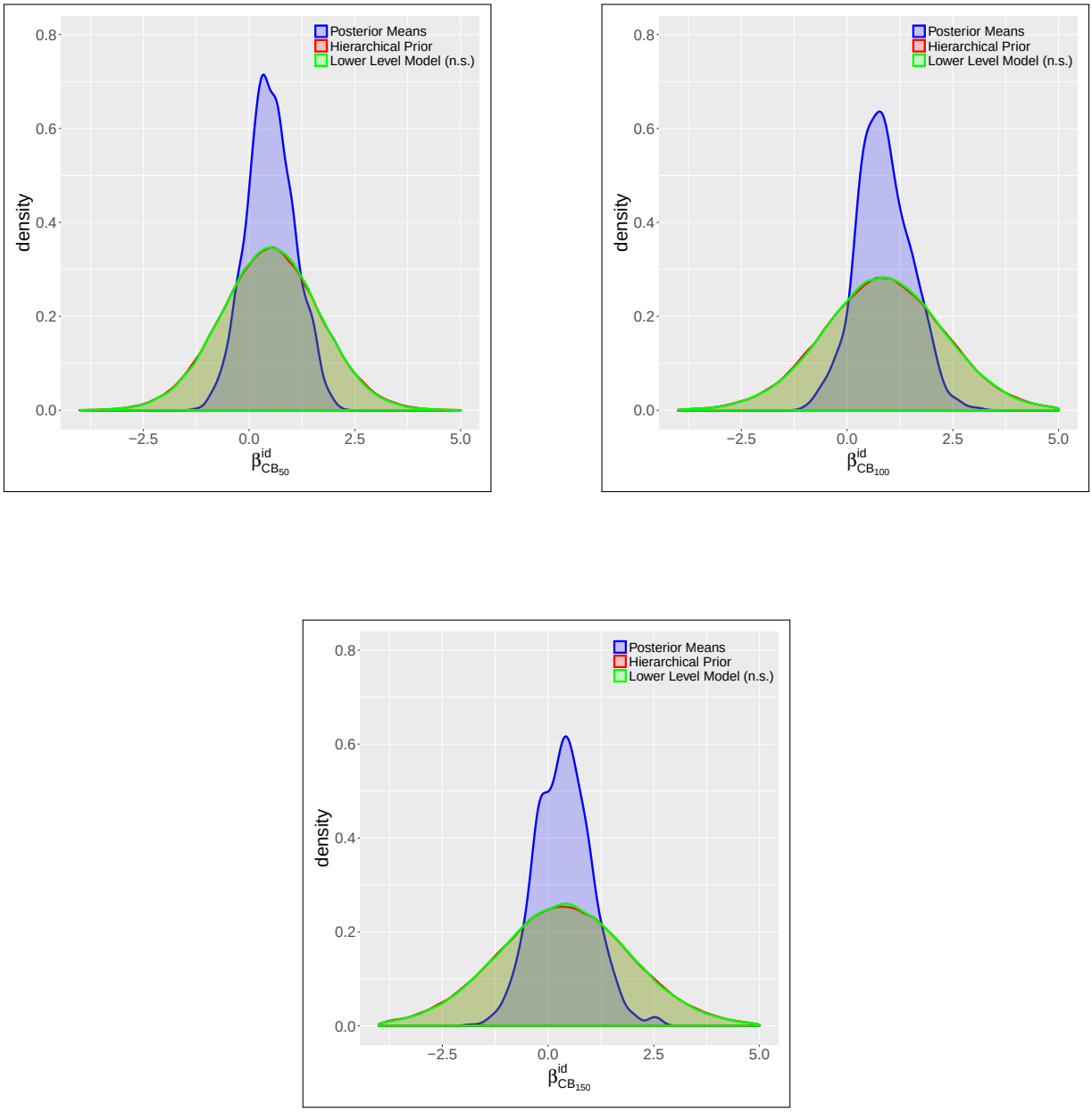

Figure 10: Posterior predictive population densities of the levels of the cash back attribute using posterior means, the posterior of the hierarchical prior and lower level model (n.s.) in an unconstrained model.

	$\beta^{id}_{CB_{50}}$	$\beta^{id}_{CB_{100}}$	$\beta^{id}_{CB_{150}}$
1%	-2.370	-2.633	-3.802
25%	-0.335	-0.106	-0.824
50%	0.481	0.884	0.325
75%	1.294	1.886	1.465
99%	3.329	4.421	4.444

Table 20: Quantiles of posterior predictive population densities of the levels of the cash back attribute in an unconstrained model.