



Journal of Marketing Research  
Article Postprint  
Volume XLV  
© 2008, American Marketing Association  
Cannot be reprinted without the express  
permission of the American Marketing Association.

## The Value of Informative Priors in Bayesian Inference with Sparse Data

Peter Lenk<sup>a</sup> and Bryan Orme<sup>b</sup>

**Acknowledgements:** We wish to thank the editor and reviewers for making helpful suggestions that greatly expanded the scope of this research. Rich Johnson also provided many insights to the analyses.

- a. Professor of Marketing and Operations Management Science, Stephen M. Ross Business School at the University of Michigan, 701 Tappan Street, Ann Arbor, MI. 48109-1234. Phone: 734-936-2619; fax: 734-936-0274; e-mail: [plenk@umich.edu](mailto:plenk@umich.edu).
- b. President, Sawtooth Software, 530 West Fir Street, Sequim, WA 98382-3209. Phone: 360-681-2300; fax: 360-681-2400; email: [bryan@sawtoothsoftware.com](mailto:bryan@sawtoothsoftware.com).

## **The Value of Informative Priors in Bayesian Inference with Sparse Data**

Informative prior that reflect the structure of the model can improve estimation when data are sparse, while “standard,” non-informative priors can have unintended consequences. The authors first discuss selecting informative priors for variances and introduce a conjugate prior for covariance matrices. The proposed prior is more flexible than the inverse Wishart without increasing computations. Second, the authors investigate the impact of priors for the covariance of parameter heterogeneity when the predictor variables are qualitative. Estimates of the omitted effects are spurious with the standard prior. The authors propose an effects-prior that treats all effects symmetrically. Third, the authors consider willingness-to-pay. These ratio estimators magnify uncertainty in the price coefficients and can give unreasonable values for price insensitive subjects. The authors show that estimation of willingness-to-pay can be greatly improved by restricting the parameters without distorting the parameters. In all three cases, the standard, non-informative priors perform very well for non-sparse data. For sparse data the proposed prior distributions better match the structures of the marketing research problems and improve inferences.

**Keywords:** choice based conjoint, variance estimation, willingness-to-pay, priors, effects-coding, hierarchical Bayes

Marketing researchers constantly push the performance envelope of methodology as clients make ever-greater demands to investigate more complex phenomenon. Drivers of this trend are models that reflect wider ranges of marketing actions, rapidly shifting and evolving customer preferences, new distribution channels and advertising media, more sophisticated users of marketing research, and cheaper computing power. Even though datasets become increasingly large with the advent of point-of-sales data and Internet data collection, data can be sparse as models become more complex. The number of observations per parameter can be quite low although the overall size of the dataset is immense. This trend is especially true for hierarchical Bayes (HB) models that capture the heterogeneity in subject-level parameters where sparse datasets are the rule and not the exception. It is not unusual to have negative degrees-of-freedom<sup>1</sup>: more parameters than observations.

To illustrate the trend in model complexity, Fader, Lattin, and Little (1992) fitted a 12 parameter, aggregate, logit model for brand choice to 3,079 purchases of frozen orange juice made by 200 households. Two years later, Allenby and Lenk (1994) fitted a hierarchical Bayes, logistic-normal regression model to 4,312 brand choices of ketchup made by 735 households. The HB model had six household-specific parameters for 4,410 parameters, excluding fixed effects and heterogeneity parameters. At the household level, there was less than one brand choice per household parameter on average. More recent articles, such as Chung and Rao (2003), Seetharaman, Ainslie and Chitagunta (1999) and Singh, Hansen and Gupta (2005), to mention only a few instances, are exemplars of this trend. In fact, Rossi, McCulloch, and Allenby's (1996) persuasive argument for using purchase history data in targeted marketing naturally results in models that often have more parameters than observations.

Sparse data occur rather routinely with choice data or conjoint experiments, the primary focus of this paper. For example, the parameters space for the choice-based conjoint (CBC) experiment reported in Arora and Huber (2001) is rather modest compared to commercial applications. Even so, their experiment resulted in six, individual-level parameters for 18 choice tasks – only three choice tasks per parameter. Marshall and Bradlow (2002) fitted 14, individual-level parameters to 24, individual-level observations – slightly more than one observation per parameter. It is not unusual for individual-level parameters to exceed the number of individual-level observations. Gilbride and Allenby (2004) report a CBC study with only 14, individual-level observations to estimate 18, individual-level parameters, excluding cut-points for their disjunctive and conjunctive models. In addition to inference, HB models provide a natural framework for predicting future observations from predictive distributions (Lenk and Rao 1990 and Lenk 1992), which are routinely used in CBC to obtain market share predictions.

Bayesian models really shine in “broad and shallow” situations where there are many sampling units and few observations per unit, such as the previous examples. In Markov chain Monte Carlo (MCMC) the distribution of individual-level parameter heterogeneity becomes a pseudo prior distribution when estimating a subject’s parameters. The parameters for this distribution are aggregate estimators of population-level parameters. Even when the data are too sparse to obtain maximum likelihood estimators, HB models stably estimate individual-level parameters by optimally pooling data across sampling units. Lenk, DeSarbo, Green and Young (1996) demonstrate this HB advantage by randomly dropping profiles from a conjoint dataset. In fact, Bayes estimators exist even without sampling data, though they merely reflect the prior assumptions of the model. With sparse data, the prior specification plays an important role in

determining the degree of aggregation from different sampling units. However, prior distributions are double-edged swords. They lend stability to ill-behaved likelihood functions, but prior assumptions that seem rather innocuous for non-sparse datasets can lead to unexpected results for sparse ones. Informative priors that better reflect salient features of the model can improve estimation in low sample information situations.

Why are priors important? Underlying the success of Bayesian estimators are appealing, theoretical properties, such as admissibility (c.f. and Bernardo and Smith 1994, DeGroot 1970 and Ferguson 1967) and coherence (De Finetti 1937). Given a loss function for estimation error or prediction error, Bayes rules (estimators) minimize expected posterior loss and are admissible: there does not exist another estimator that performs better for every value of the parameter. Coherence is a property of probability specifications and comes from gambling. A gambler cannot arrange bets to lock-in a guaranteed profit if and only if the bookie states odds in accordance to probability functions. “Coherence” is also called “arbitrage” and is a cornerstone of financial theory. Bayesian inferences and predictions with proper priors are automatically coherent, which is not true for frequentist methods. Other theoretical insights often focus on large-sample behavior. If the prior distributions put positive probability on neighborhoods of the true parameters, then Bayes estimators are consistent, and posterior distributions are asymptotically normal (c.f. Berger 1985 or Doob 1953). Of course, for large sample sizes Bayes estimators will be very close to other estimators, such as maximum likelihood or generalized method of moments, assuming equivalent models. Priors add the most value with small samples.

The above theory requires prior distributions. There are two different opinions about their selection: choose a prior that introduces as little information as possible or construct the

prior to improve estimation by leveraging non-sample information. A growing Bayesian literature is prescriptive about which priors “ought” to be used for different models. These priors are sometimes called “objective” (c.f. Berger and Sun 2008) because they remove the analyst from determining the prior. If modelers employ an agreed-upon algorithm for selecting priors and that algorithm does not elicit subjective information about the parameters, then the prior could credibly be called “non-subjective.” If all modelers use the same algorithm, then there will not be disagreement about the “correct” prior. Proper, objective priors allow the user to leverage the desirable properties of Bayesian inference while avoiding contentious issues about subjective information. These non-informative priors tend to be relatively flat over the parameter space so that the information in the sample data dominates the prior distribution with non-sparse data, and the posterior distribution is nearly identical to the likelihood function rescaled to integrate to one. In many cases, non-informative priors are improper and integrate to infinity, which negates some of the advantageous properties of Bayesian inference, such as admissibility (Stone and Dawid 1973).

Algorithms for constructing objective priors commonly impose extra-Bayesian criteria. The most famous example is Jeffrey’s (1946) invariance prior. If the analyst is ignorant about the value of a parameter, then he or she should also be ignorant about transformations of the parameter. Invariant priors are right Haar measures (c.f. Helland 2004). A limitation of most invariant priors is that they are not proper. Bernardo (1979) and (2007) and Berger and Bernardo (1989) and (1992) address this limitation with reference priors, which maximize the Kullback-Leibler divergence criterion for the expected information of an experiment. Intuitively, a reference prior should have as little impact on the posterior distribution as possible.

Jeffrey's invariance priors are sometimes special cases of reference priors. Bernardo and Ramon (1998) list five properties for producing objective priors, of which reference priors satisfy them all. Datta and Ghosh (1995) critically compare different receipts for creating objective priors. Bernardo and Smith (1994) allow moment constraints in objective priors. Kass and Wasserman (1996) provide an outstanding overview of objective priors. Objective priors tend to be difficult to specify in general, and they have not been used in HB models due to the intractability of the optimization criterion.

A different stream of literature makes positive use of priors to improve estimation by describing qualitative information, to regularize inference problems, or to impose structural constraints. Nonparametric Bayesian inference relies heavily on informative priors because the parameter space consists of functions, which are infinite dimensional parameters (Antoniak 1974, Blackwell and MacQueen 1973, and Ferguson 1973). Data are always sparse in nonparametric models. Maximum likelihood estimators do not exist or are not your everyday, garden-variety function. In density estimation, the likelihood goes to infinity as the density converges to infinitely large spikes (Dirac-delta functions) at the observations. Similarly, the regression function becomes a scatter plot. In a role reversal from objective priors, the likelihood functions are as non-informative as possible, while the prior distributions provides all of the structure for the analysis.

Craven and Wahba (1979), Good and Gaskins (1980), and Wahba (1983) propose penalized maximum likelihood to regularize estimation of nonparametric density and regression by adding a penalty term to the likelihood function. The penalty term keeps the estimators and likelihood function from going to infinity and corresponds to prior information about the desired

smoothness of the function space. Penalized maximum likelihood estimators are posterior modes, which are Bayes rule for 0/1 loss functions. The prior precision parameter determines the tradeoff between the data and the prior and is usually estimated by predictive cross validation (Stone 1974 and Geisser 1975). Evgeniou, Pontil, and Poggio (2000) provide an excellent review of related techniques that were independently developed in machine learning. Lenk (1993), (1999) and (2003) develops a fully Bayesian model with hierarchical priors for both the degree of smoothness and the tradeoff between the prior and likelihood.

Prior distributions can also be used to impose structural constraints on parameters. Gelfand, Smith, and Lee (1992) describe MCMC methods for constrained parameters. Allenby, Arora, and Ginter (1995) improve the predictive performance of discrete-choice conjoint by using self-explicated importance ratings as prior constraints on partworth parameters. Boatwright, McCulloch, and Rossi (1999) impose constraints on the sign of coefficients. These hard constraints ensure that the estimated model agrees with economic theory. The constraints do not always improve estimation, but they can greatly improve predictions by ruling out non-economic behavior.

In day-to-day practice, Bayesian marketing researchers employ standard families of prior distributions that work well with common models and flexibly encode prior information. Within these families, analysts select prior parameters. However, there may not be compelling reasons, other than convenience and tradition, for the choice of the prior family and its parameters. Provided they put positive mass on the likely region of the parameter space, the specific choices for the priors are usually not too critical, except in sparse data situations. A common prior assumption for regression parameters, such as brand preferences or advertising effects, is that

their prior mean is zero with a large prior variance. This prior assumption maintains that the independent variables are not predictive and that the researcher is not confident of his or her assessment. These proper but not very informative priors suit academic research well because they bias the results towards the null hypothesis that the model is not more predictive than chance. In addition, sensitivity studies often indicate that this specification is robust, provided that the prior variance is sufficiently large. Proper but non-informative prior distributions for variances are more problematic to specify. Unlike regression parameters where zero indicates no effect, there does not exist a single value for error variances in all settings that correspond to the null hypothesis that the model is not more predictive than chance. In regression, the equivalent null hypothesis is that the error variance is equal to the unconditional variance of the dependent variable. In addition, a seemingly non-informative specification for the variance is highly informative for its inverse, the error precision.

This paper elaborates this theme: seemingly non-informative priors can have unintended estimation properties when applied to sparse-data situations. We argue that the search for non-informative prior specifications can be counterproductive in some situations and that well constructed, informative priors can improve estimation while still being true to the data. The paper focuses on variance estimation and considers three different methods of introducing informative priors: specifying prior parameters to improve estimation; specifying a more flexible class of priors than the inverse Wishart; and adding restrictions to the model that limits estimation uncertainty.

The next section takes a closer look at the inverse Gamma distribution (IG), which is commonly used as a prior for variances, identifies some of its pitfalls, and recommends

informative priors to avoid them. Most importantly, IG distributions have a “dead zone” near zero where the density becomes very small. If the true variance is in the dead zone, it will be poorly estimated, especially with sparse data. For this reason, we believe that users need to select their prior distributions carefully, and we suggest a method for doing so. We then consider its multivariate generalization, the inverse Wishart (IW) distribution for covariance matrices. A limitation of the IW distribution is there is only one degrees-of-freedom or shape parameter for all of the variances, which implies that the coefficients of variation for the variances are equal. This limitation restricts users’ ability to encode prior information. We consider an alternative to the IW that is flexible yet is also conjugate and requires about the same amount of computation as the IW.

We then consider the covariance for partworth heterogeneity in choice-based conjoint with discrete predictors. We demonstrate that the implied estimates of the omitted level and market share predictions can appear to be spurious with sparse data when using standard prior specifications. We then provide a simple solution by using non-diagonal, prior scale matrices for the IW distribution. Unlike the standard prior specification, the proposed one treats the included and excluded effects symmetrically.

Lastly, we consider willingness-to-pay (WtoP) estimators in CBC. WtoP is estimated by dividing attributes’ coefficients by the price coefficient. WtoP is widely used for goods and services that lack market prices (Louviere, Hensher and Swait 2000 and Bennet and Blamey 2001), such as public policy, environmental economics, and healthcare. Viscusi, Magat, and Huber (1991) infer WtoP from adaptive paired comparison experiments. Mackenzie (1993) compares four measurement methods of conjoint and their impact on WtoP for attributes of

hunting trips. Train and Atherton (1995) compute WtoP for energy efficient appliances, and Ryan and Hughes (1997) estimate WtoP for pain in miscarriage management. How well does CBC WtoP work? Carlsson and Martinsson (2001) do not find significant differences between estimated WtoP and actual donations to environmental projects. However, Lloyd (2003) reviews possible sources of biases in preference elicitation, such as simplifying heuristics.

Meijer and Rouwendal (2006) and Sonnier, Ainslie, and Otter (2007) highlight issues in estimating WtoP for random coefficients models and HB models, respectively. WtoP is ratio estimator, which is notoriously difficult to estimate when the dominator is close to zero. Efron (1987) motivates bootstrap confidence intervals with ratio estimators, and Diccio, Hall, and Romano (1991) recommend Bartlett's correction to improve their sampling properties. These frequentist procedures are ill suited for HB models. Estimation uncertainty in the price coefficient becomes magnified in WtoP when the price coefficient is near zero. If subjects obey economic theory, then WtoP can be accurately estimated provided there is sufficient sample information for each subject. Unfortunately, the amount of sample information may be insufficient to do so for price insensitive subjects. We propose bounding the price coefficient away from zero, and limiting the size of the estimated WtoP measures by using parameter restrictions. These constraints on regression coefficients indirectly impose constraints on estimation variances. The constraints improve estimation of the WtoP without distorting the estimated partworths.

## *INFORMATIVE PRIORS FOR VARIANCES AND COVARIANCE MATRICES*

### *Specifying Prior Distributions for Variances*

Proper but non-informative prior distributions for variances and covariance matrices can be difficult to specify. Jeffrey's (1946) non-informative prior for the variance  $\sigma^2$  of a normal distribution is proportional to  $\sigma^{-2}$ . Despite appealing theoretical properties, such as being invariant to scale transformations of the observations, it is very unreasonable in practice. For instance, the probability that  $\sigma^2$  is between 1 and 2 is the same as between 1000 and 2000 and between .01 and .02. Surely, marketing researchers can make better guesses than those implied by Jeffrey's prior. Unfortunately, there does not exist one proper but non-informative prior that works well in all settings, which forces users to employ informative priors that depend on the context.

We advise the use of informative prior distributions for the variance in sparse data situations. Researchers should roughly know the scale of their measurements, at least to an order of magnitude. A useful method for developing a guess for the standard deviation  $\sigma$  is to consider typical high and low values (say 2.5<sup>th</sup> and 97.5<sup>th</sup> percentiles), and divide their difference by four. This procedure over-estimates the standard deviation of leptokurtic distributions, but usually not so much to adversely affect Bayesian inference. For instance, ratings on a seven-point scale often have standard deviations around 1.5 or less, and the theoretical, maximum standard deviation<sup>2</sup> is 3. It would be unreasonable to specify a prior that has much mass above the maximum. Next, the user specifies an uncertainty measure, such as a standard deviation or percentiles. Based on the most likely value and the measure of uncertainty, one can back out the parameters for informative prior distributions.

Bayesians often use the inverse Gamma (IG) distribution as a prior for variances. The density of IG( $\alpha/2, \beta/2$ ) distribution with  $\alpha > 0$  and  $\beta > 0$  for  $\sigma^2$  is:

$$f(\sigma^2) = \frac{\left(\frac{\beta}{2}\right)^{\frac{\alpha}{2}}}{\Gamma\left(\frac{\alpha}{2}\right)} (\sigma^2)^{-\left(\frac{\alpha}{2}+1\right)} \exp\left(-\frac{\beta}{2\sigma^2}\right) \text{ for } \sigma^2 > 0. \quad (1)$$

The shape parameter or degrees-of-freedom  $\alpha$  is often called the “prior sample size.” Smaller values of  $\alpha$  correspond to less-informative priors. The scale parameter is  $\beta$ . The mode, mean and variance are:

$$\text{Mode} = \frac{\beta}{\alpha + 2}; E(\sigma^2) = \frac{\beta}{\alpha - 2} \text{ for } \alpha > 2; \text{ and } V(\sigma^2) = \frac{2}{\alpha - 4} E(\sigma^2)^2 \text{ for } \alpha > 4. \quad (2)$$

The coefficient of variation (CV) is  $[2/(\alpha-4)]^{1/2}$ , so the larger the prior degrees-of-freedom, the more informative the distribution.

[Figure 1]

Figure 1 plots the IG for 2, 4, 6, 8, and 10 degrees-of-freedom with the mode set to 10. The IG has two properties – a “dead zone” and very long tails - that are not well appreciated. Although, the support of the distribution is the positive real numbers, there is an interval close to zero where the IG density is essentially 0. When  $\alpha = 10$  and  $\beta = 120$ , the density drastically decreases before 3:  $f(3) \approx 10^{-4}$ ;  $f(2) \approx 10^{-8}$ ; and  $f(1) \approx 10^{-19}$ . Variances in the dead zone will require unexpectedly large sample sizes to be estimated accurately. After the dead zone, the IG rapidly increases to the mode. The skew of the inverse Gamma distribution is  $4(2\alpha-8)^{1/2}/(\alpha-6)$  for  $\alpha > 6$  and the right tail declines algebraically at rate  $(\sigma^2)^{-(\alpha+2)/2}$ . The tail is very long when  $\alpha$  is less than 2. In Figure 1, the 95<sup>th</sup> percentile is 390 when  $\alpha = 2$  and  $\beta = 40$ . The long-tailed property for small  $\alpha$  implies that users need not worry too much about estimating large variances

if the mode has the correct order of magnitude. Small variances are more problematic if they fall into the dead zone.

One method to specify an informative IG prior is to give the mean  $m$  and standard deviation  $s$  for  $\sigma^2$  and solve for the IG parameters:  $\alpha = 2(m/s)^2 + 4$  and  $\beta = m(\alpha - 2)$ . One limitation is that  $\alpha$  will always be larger than 4. Increasing the standard deviation while holding the mean fixed will not have much impact on the IG density once  $\alpha$  is less than 5. An alternative method is to specify the mode and a percentile, say 90<sup>th</sup>. The percentile for the inverse Gamma distribution can be computed from the cumulative distribution function  $G$  for the Gamma distribution, which is widely available. The relationship is  $p = P(\sigma^2 < x) = 1 - G(1/x)$  where  $x$  is the 100p percentile for the inverse Gamma distribution. One method in EXCEL is to use Solver to find the  $\alpha$  that minimizes:

$$h(\alpha) = [1 - \text{GAMMADIST}(1/x, \alpha/2, 2/\{M(\alpha+2)\}) - p]^2 \text{ and set } \beta = M(\alpha+2). \quad (3)$$

If the user has a good idea about the maximum variance, then it would make sense to use a truncated IG prior. Truncated priors are easily handled in hierarchical Bayesian models if they are at the top of the hierarchy by using the inverse cumulative distribution method (Gelfand, Smith, and Lee 1992). Random draws are generated from the truncated distribution using the inverse cdf:  $\sigma^2 = F^{-1}[uF(\text{MaxV})]$  where  $F$  is the cumulative distribution function for the inverse Gamma distribution;  $u$  is a uniform random number, and  $\text{MaxV}$  is the maximum variance. The next section implements this procedure for covariance matrices.

*Specifying Prior Distributions for Covariance Matrices*

In this section we develop an alternative to the inverse Wishart (IW) (c.f. Zellner 1971) distribution for the covariance matrix of a multivariate normal distribution. The Wishart and IW are the multivariate generalization of the Gamma and IG distributions. The marginal distribution of each variance component of an IW distribution has an IG distribution. A limitation of the IW distribution is that the degrees-of-freedom parameters are equal for all of the variances. Consequently, the variances' coefficients of variation are equal, which may not adequately describe prior beliefs. Our alternative distribution allows different degrees-of-freedom parameters for the variances. In addition, it is also conjugate to the normal distribution, and does not require Metropolis algorithms.

To fix ideas,  $\underline{Y} = (Y_1, \dots, Y_m)'$ , a  $m$ -vector of random variables, has a multivariate normal distribution:  $\underline{Y} \sim N_m(\mu, \Sigma)$ . Without loss of generality, the mean vector  $\mu$  can be a function of covariates. The covariance matrix is  $\Sigma = [\sigma_{ij}]$  for  $i, j = 1, \dots, m$ , and  $\Sigma_{k,k}$  is a sub-matrix of  $\Sigma$  consisting of its first  $k$  rows and columns. Partition  $\Sigma_{k,k}$  as:

$$\Sigma_{k,k} = \begin{bmatrix} \Sigma_{k-1,k-1} & \Sigma_{k-1,k} \\ \Sigma_{k,k-1} & \sigma_{k,k} \end{bmatrix} \quad (4)$$

where  $\Sigma_{k-1,k}$  is a  $k-1$  vector of covariances between  $(Y_1, \dots, Y_{k-1})'$  and  $Y_k$ , and  $\Sigma_{k,k-1}$  is the transpose of  $\Sigma_{k-1,k}$ . When  $k=1$ ,  $\Sigma_{1,1} = \sigma_{1,1}$ , and  $\Sigma_{0,1}$  is undefined.

Instead of working with the  $\Sigma$  directly, we reparameterize it as follows:

$$\tau_1^2 = \sigma_{11}; \tau_k^2 = \sigma_{kk} - \Sigma_{k,k-1} \Sigma_{k-1,k-1}^{-1} \Sigma_{k-1,k}; \text{ and } \gamma_k = \Sigma_{k-1,k-1}^{-1} \Sigma_{k-1,k} \text{ for } k = 2, \dots, m. \quad (5)$$

These parameters are used in the well-known definition of the means and variances for conditional normal distributions. If one knows  $\{\tau_k^2\}$  and  $\{\gamma_k\}$ , then the elements of  $\Sigma$  can be obtained through recursion. Start the recursion with  $\sigma_{11} = \tau_1^2$  and  $\sigma_{12} = \sigma_{11}\gamma_2$ . Next, suppose

that the elements of  $\Sigma_{k-1,k-1}$  are known. Then the recursion for  $k$  is  $\Sigma_{k-1,k} = \Sigma_{k-1,k-1} \gamma_k$  and  $\sigma_{kk} = \tau_k^2 + \gamma_k' \Sigma_{k-1,k-1} \gamma_k$ .

The joint density of  $\underline{Y}$  is the product of univariate, conditional normal densities:

$$f(\underline{y}) = f_1(y_1) \prod_{k=2}^m f_{k|1\dots k-1}(y_k | y_1, \dots, y_{k-1}) \quad (6)$$

where

$$Y_1 \sim N(\mu_1, \sigma_{11}) \text{ and } Y_k | Y_1 \dots Y_{k-1} \sim N(\mu_k + \gamma_k' R_{k-1}, \tau_k^2) \text{ for } k = 2, \dots, m, \quad (7)$$

and  $R_k = (Y_1 - \mu_1, \dots, Y_k - \mu_k)'$  are the residuals. The  $\gamma_k$  are the regression coefficients for the vector of residuals  $R_{k-1}$ , and  $\tau_k$  is the conditional standard deviation. We induce a prior on  $\Sigma$  by defining distributions on  $\{\tau_k^2\}$  and  $\{\gamma_k\}$ . We will assume that these parameters are mutually independent with the following distributions:

$$\tau_k^2 \sim IG\left(\frac{\alpha_k}{2}, \frac{\beta_k}{2}\right) \chi(\tau_k^2 \leq T_k^2) \text{ for } k = 1, \dots, m \text{ and } \gamma_k \sim N_{k-1}(\lambda_k, \Psi_k) \text{ for } k = 2, \dots, m \quad (8)$$

where  $\chi$  is the indicator function, and  $T_k^2$  is the known upper bound for  $\tau_k^2$ ;  $T_k^2$  could be infinity. Jen, Chou, and Allenby (2008) use a similar approach for bivariate normals. We will term this prior the ‘‘Conditional Normal/Inverse Gamma’’ distribution (CNIG). The IW distribution has  $(m)(m+1)/2$  scale parameters and one, degrees-of-freedom parameter. In comparison, the  $\{\tau_k^2\}$  has  $m$  degrees-of-freedom parameters and  $m$  scale parameters, and  $\{\gamma_k\}$  has  $(m)(m-1)/2$  mean parameters and  $(m)(m-1)(m-2)/6$  variance and covariance parameters, assuming independent  $\{\gamma_k\}$ . If properly utilized, these additional parameters enhance flexibility in the prior specification. The full conditional distributions for  $\{\tau_k^2\}$  and  $\{\gamma_k\}$  are also IG and

normal. Details are in Web Appendix A. The CNIG distribution and IW distribution, using Bartlett's decomposition (c.f. Ripley 1987), have approximately the same simulation effort.

A brief simulation study proves the concept of the CNIG prior. The simulation model was a multivariate regression with  $m=4$  dimensions for the dependent variable and two covariates. The error variances are .1, 1, 10, and 100. We selected this specification to mimic the situation where the scales of the dependent variables are much different. We considered a sparse data situation with sample size  $n = 10$  to estimate 12 regression parameters and 10 variance and covariance parameters. We use a standard IW prior where the prior mean is the identity and the prior degrees-of-freedom is six:  $2 + m$ . For the CNIG we use informative priors for  $\{\tau_k^2\}$  by specifying the mode, 90<sup>th</sup> percentile, and maximum in the top-right side of Table 1. The  $\{\gamma_k\}$  have independent normal distributions with mean 0 and standard deviation 10.

Table 1 displays the true  $\Sigma$  and its posterior means and standard deviation under the IW and CNIG priors. The CNIG prior gives better estimates than the IW prior because the former includes more prior information. Not surprisingly given the small amount of sample information, the posterior standard deviations are fairly large. When the sample size is increased to  $n = 100$  (simulation results not reported here), both models become very accurate, but the CNIG retains a slight advantage.

[Table 1]

The CNIG prior allows greater flexibility to encode prior information, particularly prior uncertainty, than the IW. When the sample information dominates prior information, the two prior models give very similar answers. The CNIG can be useful with sparse data when the prior coefficients of variation are different. A practical, common alternative is to standardize the data

to a common scale. However, the error variance can be different depending on the strength of the predictor variables. This procedure does not imply that the variances have the same prior CVs, which is the assumption of the IW. To be technically correct, the standardization should result in equal CVs. Web Appendix A further discusses implementation issues.

### *EFFECTS PRIORS FOR COVARIANCES*

The canonical model that motivates this section is choice-based conjoint (CBC) studies where the HB model has two levels (c.f. Allenby, Arora, and Ginter 1998; Allenby and Ginter 1995, and Huber, Wittink, Fielder, and Miller 1993). The within-subject model specifies subject  $i$ 's latent utility latent  $U_{i vw}$  for profile  $w$  in choice task  $v$  as:  $U_{i vw} = \mathbf{x}'_{i vw} \beta_i + \varepsilon_{i vw}$  where  $\mathbf{x}_{i vw}$  is the design vector for this profile;  $\beta_i$  is a vector of parameters specific to subject  $i$ ; and  $\varepsilon_{i vw}$  is a random error term, either a multivariate normal distribution for probit models or extreme value distribution for logit models. The between-subjects model captures the heterogeneity across subjects of the individual level parameters  $\beta_i$ . One of the simplest specifications assumes that they are a random sample from a multivariate normal distribution  $\beta_i = \theta + \delta_i$  where  $\theta$  is the population mean of the subject-specific parameters;  $\delta_i$  is a multivariate normal error term with mean zero and covariance matrix  $\Delta_\beta$ . Prior distributions are specified for  $\theta$  and  $\Delta_\beta$ . This setup can be generalized to other forms of heterogeneity. One such generalization is a multivariate regression models for the individual-level parameters (c.f. Lenk, DeSarbo, Green, and Young 1996 and Rossi, McCulloch, and Allenby 1996) and mixture of normal components (c.f. Allenby, Arora, and Ginter 1998 and Lenk and DeSarbo 2000). The focus of this section is the prior specification for the covariance matrix  $\Delta_\beta$  for the unexplained heterogeneity in the

individual-level parameters, and the exact form of the between-subjects model is not important to the paper's thesis.

### *Evidence of a Problem with Sparse Datasets*

We illustrate the impact of an unfortunate choice of prior distributions with a simulated dataset. The simulated data were generated to mimic a CBC experiment. There is one attribute (“brands”) that has 12 levels. Effects coding was employed for the brand preferences where brand 12 was omitted. For  $j = 1$  to 11 the effects variables are:  $x_j = 1$  for brand  $j$ ,  $x_j = -1$  for brand 12, and  $x_j = 0$  otherwise. The random utility for subject  $i$  and brand  $j$  is  $U_{ij} = \beta_{ij} + \varepsilon_{ij}$  for  $j = 1, \dots, 11$ , and  $U_{i,12} = -\beta_{i1} - \dots - \beta_{i,11} + \varepsilon_{ij}$  where  $\varepsilon_{ij}$  is the error term from an extreme value distribution. Each choice task consists of six alternatives, which were randomly selected without replacement from the 12 brands. The number of “subjects” is 300. The true value for all subject-level parameters is zero. The HB model assumes multivariate normal heterogeneity:  $\beta_i \sim N_{11}(\theta, \Delta_\beta)$  with priors distributions  $\theta \sim N_{11}(0, 10 * I)$  and  $\Delta_\beta^{-1}$  is Wishart with prior mean equal to the identity matrix and prior degrees-of-freedom equal to 16.

[Table 2]

Table 2 presents the posterior means and standard deviations of the brand effects as the number of choice tasks per subject varies from two to 20. When there are two choice tasks per subject, the posterior mean ranges from  $-.08$  to  $.49$  for brands one to 11. The omitted brand, brand 12, appears to be an outlier at  $-2.14$ , which would lead one to conclude that brand 12 is very undesirable on the logit scale. In actuality, all of the true parameters were zero. As the number of choice tasks increases per subject, all of the coefficients tend to zero; however, the omitted brand remains an outlier until 10 choice tasks.

[Table 3]

Because the posterior means are less, in absolute value, than the posterior standard deviations, one should conclude that the brand effects, averaged across subjects, are zero; though brand 12 has an unusually large standard deviation. Should one worry about the inflation in the standard deviation for brand 12? Table 3 presents simulated choice shares for the 12 brands using the estimates in Table 2 where the  $\{\beta_i\}$  were drawn from the predictive distribution of heterogeneity. The true choice shares are 8.33%. With two choice tasks per subject, the choice share for brand 12 is grossly overstated, and its standard deviation is large when compared to the other brands. This inflation of brand 12's choice share is due to the large variance for its effect. The simulator frequently draws values for brand 12 that exceed the other brands. This effect attenuates as the number of choice tasks per subject increases.

#### *Effects Prior for One Attribute*

We diagnosis the source of the problem and propose a solution and use standard terminology from experimental design and ANOVA. The simulation study in the previous section is an example of a one-way layout where there is one attribute or factor with  $K$  levels or treatments. The basic issue is that transformations of the treatment means may be at odds with the default prior covariance specification. The treatment means are  $\mu = (\mu_1, \dots, \mu_K)'$ . The ANOVA parameterization uses the grand mean and treatment effects:

$$\bar{\mu}_\bullet = \frac{1}{K} \sum_{k=1}^K \mu_k \text{ and } \alpha_k = \mu_k - \bar{\mu}_\bullet \text{ for } k = 1, \dots, K. \quad (9)$$

The effects sum to zero, and the model is not identified if all  $K$  effects are included in the model. One of the effects, say  $\alpha_K$  is excluded, and the omitted effect is defined as the negative of the sum of the included effects. The effects that are included in the model are  $\alpha = (\alpha_1, \dots, \alpha_{K-1})'$  and

the omitted effect is  $\alpha_K = -(\alpha_1 + \dots + \alpha_{K-1})$ . With choice data, the grand mean, which is also the model's intercept, is set to zero to identify the latent utilities in choice-based conjoint.

If the default prior covariance for the included effects  $\alpha$  is  $v^2 I_{K-1}$ , a constant times the identity matrix, then the implied prior variance for the omitted effect  $\alpha_K$  is  $(K-1)v^2$ , and the covariance between the omitted and included effects are  $-v^2$ . The prior variance for the omitted effect is  $K-1$  times larger than the included effects. This additional uncertainty for the omitted effect leads to greater variation in its estimator, especially for small sample sizes. The default prior treats the included and excluded treatment effects asymmetrically, which is usually not the researchers' intent. By considering the inverse transformation, the induced prior covariance on the treatment means is also asymmetric:

$$\text{cov}(\mu) = \Sigma_{\mu} = v^2 \begin{bmatrix} I_{K-1} + J_{K-1} & \underline{0}_{K-1} \\ \underline{0}'_{K-1} & K+1 \end{bmatrix} \quad (10)$$

where  $I_{K-1}$  is the  $K-1$  dimensional identity matrix;  $J_{K-1}$  is a  $(K-1)$  by  $(K-1)$  matrix of ones; and  $\underline{0}_{K-1}$  is a  $K-1$  vector of zeroes. This implied prior specification, which gives the last treatment a prior variance that is  $(K+1)/2$  times greater than the other variances, is probably not the one that the research had in mind when specifying a non-informative prior on the model's parameters.

A remedy for this situation is to assume that the default prior applies to the treatment means and derive the implied prior for the treatment effects. Suppose that the prior covariance of the treatment means  $\mu$  is  $\sigma_{\mu}^2 I_K$ . Then the induced prior covariance of the  $K-1$  included effects is symmetric:

$$\text{var}(\alpha_j) = \sigma_{\mu}^2 (K-1)/K \text{ and } \text{cov}(\alpha_j, \alpha_k) = -\sigma_{\mu}^2 / K \text{ for } j \neq k \quad (11)$$

Also, the variance of the omitted effects is  $\sigma_\mu^2(K-1)/K$ , and the covariances between the included and omitted effects is  $-\sigma_\mu^2/K$ . We will call the IW distribution with scale parameter in Equation (11) the “effects prior.” By introducing a small amount of correlation in the prior covariance, we treat the omitted and included effects symmetrically, and avoid the abnormality that the Bayes estimator of the omitted effect is discrepant from the included effects.

### *Empirical Examples*

Tables 4 and 5 continue the simulation study while using the effects covariance matrix proportional to Equation (11). Recall that Brand 12 is omitted from the model. Unlike the results for the default prior in Tables 2 and 3, Brand 12 is no longer an outlier. Its posterior standard deviation is in line with that of the other brands, and its choice share is similar to the other brands. As predicted by the analysis, the prior that symmetrically treats all of the effects corrects for the inflation of the omitted effect’s variance.

[Tables 4 and 5]

Next, we illustrate this behavior with an actual choice-based conjoint experiment. The study consisted of 1202 subjects, each of whom responded to 12 choice tasks. Each choice task used five profiles. The study included nine brands and two other attributes with eight and five levels, for 19 subject-level parameters. Unlike the simulation study, we do not know the true partworths. Consequently, to demonstrate variance inflation for the omitted levels, we ran the analysis two times where the ninth and fourth brands were omitted, respectively. Table 6 presents the estimated population-level brand effects using two tasks per subject. In both runs, the size of the effect for the omitted brand is much smaller than its size when it is included in the

model. That is, when brand 9 is omitted, its posterior expectation is  $-1.24$  less than when brand 4 is omitted. Likewise, when brand 4 is omitted, its posterior expectation is  $-1.02$  less than when brand 4 is included in the model. This empirical analysis illustrates that the default prior treats the omitted effect differently from the effects that are included in the model.

[Table 6]

Figure 2 plots the differences in the omitted and included effects. The left-hand-side uses the default, independence prior for the covariance matrix, while the right-hand-side uses the effects prior. The top panels consider the differences in posterior means, and the bottom panels consider the difference in posterior standard deviation. The graph indicates that if the number of tasks per subject is small, then the results for the default prior depends heavily on the choice of omitted level, while the result for the effects prior are invariant to the omitted level.

[Figure 2]

The empirical example demonstrates that the choice of omitted level affects the analysis when using the default prior. The default prior, instead of being nearly non-informative, actually introduces strong assumptions about the heterogeneity in the parameters. In contrast, the effects-prior corrects this situation by treating all of the effects symmetrically. Web Appendix B extends the analysis from the one-factor, main-effects model to multiple factors with interaction, and Web Appendix C repeats the analysis for dummy variables where the affect of the standard prior are less pronounced.

Consider a conjoint study that includes price where subject's  $i$  indirect utility for profile  $k$  is  $U_{ik} = x'_{ka} \beta_{ia} + x_{kp} \beta_{ip} + \varepsilon_{ik}$  where  $\beta_{ia}$  is a vector of coefficients for the non-price attributes  $x_{ka}$ , and  $\beta_{ip}$  is the coefficient for prices  $x_{kp}$ . We assume that the true price coefficients agree with economic theory and are negative; theories concerning positive price coefficients are beyond our immediate objectives. WtoP for attribute  $j$  is  $\psi_{ij} = -\beta_{iaj}/\beta_{ip}$ . Being a ratio estimate, the uncertainty in the estimates of  $\psi_{ij}$  magnifies the uncertainty in  $\beta_{ip}$  when  $\beta_{ip}$  is close to zero. To mitigate this affect, we impose constraints on the price coefficients:  $\beta_{ip} \leq c$  for  $c < 0$  and WtoP:  $|\psi_{ij}| < w_j$  for  $w_j > 0$ . Given the price coefficient, the second constraint implies  $-\beta_{ip}w_j < \beta_{iaj} < |\beta_{ip}|w_j$  for attribute  $j$ . These constraints bound the price coefficient way from zero, and keeps willingness-to-pay within a bounded range.

One approach to enforce these bounds is to use truncated normal distributions for partworth heterogeneity. Unlike the truncated IG prior, truncated distributions are difficult to implement in MCMC when they are in a middle layer of the hierarchical model. Their normalizing constants depend on the heterogeneity parameters, which cannot be ignored in their full conditional distributions. An alternative method is to consider appropriate transformations of multivariate normal random variables. We consider two transformations: the Tobit and linexp. They both introduce latent, normal random variables  $\alpha_i = (\alpha_{ia}', \alpha_{ip})'$  that have multivariate normal distributions.

The negative Tobit transformation for the price partworth for  $c < 0$  is:

$$\beta_{ip} = T_{NT}(\alpha_{ip}) = \begin{cases} \alpha_{ip} & \text{if } \alpha_{ip} \leq c \\ c & \text{if } \alpha_{ip} > c \end{cases} \quad (12)$$

The double Tobit transformation for the non-price partworths for  $w_j > 0$  and  $\beta_{ip} < 0$  is:

$$\beta_{iaj} = T_{DT}(\alpha_{iaj}) = \begin{cases} \beta_{ip} w_j & \text{if } \alpha_{iaj} < \beta_{ip} w_j \\ \alpha_{iaj} & \text{if } \beta_{ip} w_j \leq \alpha_{iaj} \leq -\beta_{ip} w_j \\ -\beta_{ip} w_j & \text{if } \alpha_{iaj} > -\beta_{ip} w_j \end{cases} . \quad (13)$$

The HB model imposes the subject-level heterogeneity on  $\{\alpha_i\}$ , not the  $\{\beta_i\}$ . The MCMC algorithm is easily implemented. See Web Appendix D for more details.

We illustrate the impact of the Tobit model in an HB Logit model with a simulation study. The simulation has four “Brands,” “Quality” and “Price.” Brand 4 is the base brand in the logit model. We generated “300” subject who respond to 30 choice tasks. Each choice task consists of four profiles with unique “brands.” The  $\{\alpha_i\}$  were generated from a multivariate normal distribution. The mean of  $\alpha_{ip}$ , the latent price coefficient, was  $-1.5$  with standard deviation of 1. The upper bound  $c$  for the price coefficient is  $-1$ , and the bound  $w$  for willingness-to-pay is 5. The truncation for the price coefficient affects around 5% of the price coefficients.

[Table 7 and Figure 3]

Table 7 details the correlations and root mean squared errors (RMSE) between the true  $\{\beta_i\}$  and their HB Logit estimates with the Tobit truncation and unconstrained (Free) estimation. The fit statistics are nearly identical for the partworths, but the Tobit estimates of WtoP are vastly superior to the Free estimates. Figure 3 plots the true versus estimated parameters with the Bayes estimates on the X axis and the true parameters on the Y axis. The graphs have  $45^\circ$  reference lines. There is relatively little to distinguish the Tobit and Free estimates of the non-price coefficients in Panels A and B. Panels C and D compare the price coefficients. The biggest difference is the due to the truncation at  $-1$ , which is only evident in the upper right hand corner of Panel D. In contrast, the effect of the Tobit truncation is more pronounced in Panels E

and F for WtoP. The Free estimates result in some wildly inaccurate estimators, with a range of -40 to 50, while the true values are between  $\pm 5$ . The practical tension is the extent to which the marketing manager believes it is feasible to restrict price sensitivity. After all, price insensitive customers make for good prospects, yet targeting and positioning should not be driven by estimation uncertainty.

The Tobit model has a limitation that could adversely affect marketing decision-making. Marketing managers often wish to compare price coefficients and WtoP among subjects for direct marketing or targeting. The Tobit models lumps together all subjects that have their latent variables outside the allowable range and does not distinguish among them. Instead of “hard” cutoffs that limit the ability to sort customers, the linexp transformation allows for “soft” censoring. The linexp is a spline function consisting of linear and exponential functions. The spline function is designed to be continuous at the knots with the same left and right hand derivatives. It is similar to the Tobit in that both are linear in the latent variable over most of the parameter space. Instead of the truncation at the boundary, the linexp uses an exponential function to asymptote to the boundary. In this way, the linexp transformation is monotonic, unlike the Tobit. The negative linexp function for the price coefficient for  $b < c < 0$  is:

$$\beta_{ip} = T_{NLX}(\alpha_{ip}) = \begin{cases} \alpha_{ip} & \text{for } \alpha_{ip} \leq b \\ c + (b - c) \exp\left[\frac{\alpha_{ip} - b}{b - c}\right] & \text{for } \alpha_{ip} > b \end{cases} \quad (14)$$

It is designed so that the left and right derivatives at the knot  $b$  are one. Similarly, the double linexp function for non-price coefficients is a “S” shaped curve between  $-w_j |\beta_{ip}|$  and  $w_j |\beta_{ip}|$ . See Equation (D.2) in Web Appendix D, which also compares the Tobit, linexp, and exponential transformations. The last two columns of Table 7 repeat the simulation study. The results are

broadly similar to those for the Tobit. The `linexp` improves estimation of WtoP without distorting the estimated partworts. The `linexp` and Tobit have nearly identical estimation accuracy.

### *CONCLUSION*

It is not at all surprising that prior distributions are important for sparse datasets that have a low observation per parameter ratio. This situation is very common in hierarchical Bayes models when the number of observations per unit or subject is small relative to the number of individual-level parameters. We analyzed three, important cases where standard or default priors may not be appropriate when applied to sparse data.

First, we argued that the quest for non-informative priors for variances is futile and recommend that practitioners use informative ones. We highlighted two important properties of the inverse Gamma distribution, which should be considered when making informative prior specifications: long tails and a dead zone around zero where the density becomes extremely small. We suggested procedures for selecting the parameters of the IG to encode prior information about the variance. Then, we considered covariance matrices and suggested a more flexible alternative to the inverse Wishart distribution, which has only one degrees-of-freedom parameter for all variances. Our alternative prior has separate degrees-of-freedom parameters for each variance, which provides more flexibility in encoding prior information for multiple variances. In addition, our alternative is conjugate with respect to normal distributions, and does not require additional computation.

Second, we consider choice-based conjoint with qualitative attributes and effects coding. The standard prior specification for the heterogeneity covariance treats the included and

excluded effects asymmetrically, which can result in unexpected and misleading results with sparse data. These anomalies are not innocuous: they can lead to misguided marketing actions or can cause clients to question studies due to insufficient face-validity. We proposed a simple solution that symmetrically treats all of model parameters for both main effects and interactions. The effects-coding prior is particularly beneficial because it eliminates the undesirable consequences of the default prior with sparse data.

Third, we examined willingness-to-pay, which is a ratio estimator. When the denominator, the subject's price coefficient, is close to zero, sampling uncertainty can lead to spurious willingness-to-pay estimates. This situation is particularly likely to occur with sparse data. We propose a model that bounds the price coefficient away from zero and limits the willingness-to-pay parameter. The bounds are enforced by using Tobit and  $\text{linexp}$  transformations. The Tobit model performs hard censoring if the latent variable falls outside of the allowable range, while the  $\text{linexp}$  transform gives soft censoring where the bounds are the asymptotes. Both methods give very similar estimation results. However, the  $\text{linexp}$  function allows the ordering of all subjects, while the Tobit model does not distinguish among subjects that violate the bounds.

Bayesian inference is not prescriptive about the choice of prior distributions: they depend on the subjective belief of the observer. Often, in an attempt to be "fair and unbiased" users select proper but non-informative priors. Sometimes these choices are actually more informative than anticipated and produce unexpected results, especially in the sparse-data situation. The conclusion of our cautionary tale is that researchers should check that their prior distributions do not have unintended consequences on the focal parameters and marketing decisions. This sage

advice is easy to offer but much harder to implement in complex models. In the simple but important cases of the paper, we were able to tackle the problem analytically. In more complex settings, a useful strategy is to simulate the parameters from the prior distribution and run through the marketing decision analysis without the benefit of data. If the implied prior distributions of the marketing action or focal parameters are not what one had in mind, e.g. too sharp, too diffuse, or in the wrong location, then one should reconsider the prior specification, especially if the data will be sparse. This analysis has to be done before obtaining the posterior distribution for the actual data: fine-tuning prior distributions to obtain desired posterior distributions invalidates Bayesian inference. Constructing priors that accurately reflect knowledge about both the model parameters and the marketing action can be challenging and is a topic for continuing research.

## REFERENCES

- Allenby, Greg M., Neeraj Arora, and Jim L. Ginter (1995), "Incorporating Prior Knowledge into the Analysis of Conjoint Studies," *Journal of Marketing Research*, 32 (May), 152–62.
- , ———, and ——— (1998), "On the Heterogeneity of Demand," *Journal of Marketing Research*, 35, (August), 384–89.
- and Jim L. Ginter (1995), "Using Extremes to Design Products and Segment Markets," *Journal of Marketing Research*, 32 (November), 392–403.
- and Peter Lenk (1994) "Modeling Household Purchase Behavior with Logistic Normal Regression," *Journal of the American Statistical Association*, 83 (428), 1218–31.
- Antoniak, Charles E. (1974), "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *Annals of Statistics*, 2, 1152–74.
- Arora, Neeraj and Joel Huber (2001), "Improving Parameter Estimates and Model Prediction by Aggregate Customization in Choice Experiments," *Journal of Consumer Research*, 28 (2), 273–83.
- Bennet, Jeff and Russell Blamey (2001), *The Choice Modelling Approach to Environmental Valuation*. Edward Elgar, Cheltenham.
- Berger, James O. (1985), *Statistical Decision Theory and Bayesian Analysis*, Second Edition. Springer-Verlag, New York.
- and Dongchu Sun (2008), "Objective Priors for the Bivariate Normal Model," *Annals of Statistics*, 36 (2), 963–82.
- and José M. Bernardo (1989), "Estimating a Product of Means – Bayesian Analysis with Reference Priors," *Journal of the American Statistical Association*, 84 (405), 200–207.

- and ——— (1992), “Ordered Group Reference Priors with Application to the Multinomial Problem,” *Biometrika*, 79 (1), 25–37.
- Bernardo, José M. (1979), “Reference Posterior Distributions for Bayesian-Inference,” *Journal of the Royal Statistical Society, Series B* (41), 113–47.
- (2007), “Objective Bayesian Point and Region Estimation in Location-Scale Models,” *SORT- Statistics and Operations Research Transactions*, 31, 3–44.
- and José M. Ramón (1998), “An Introduction to Bayesian Reference Analysis: Inference on the Ratio of Multinomial Parameters,” *Journal of the Royal Statistical Society, Series D - The Statistician*, 47 (1), 101–135.
- and Adrian F. M. Smith (1994), *Bayesian Theory*. West Sussex, England: John Wiley & Sons, Ltd.
- Blackwell, David and James B. MacQueen (1973), “Ferguson Distributions via Polya Urn Schemes,” *Annals of Statistics*, 1, 353–55.
- Boatwright, Peter, Robert E. McCulloch, and Peter 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), 1063–73.
- Carlsson, Fredrik and Peter Martinsson (2001), “Do Hypothetical and Actual Marginal Willingness to Pay Differ in Choice Experiments?” *Journal of Environmental Economics Management*, 41 (2), 179–92.
- Chung, Jaihak and Vithala R. Rao (2003), “A General Choice Model for Bundles with Multiple-Category Products: Applications to Market Segmentation and Optimal Pricing for Bundles,” *Journal of Marketing Research*, 40 (May), 115–30.

- Craven, Peter and Grace Wahba (1979), "Smoothing Noisy Data with Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 377–403.
- Datta, Gauri Sankar and Malay Ghosh (1995), "Some Remarks on Noninformative Priors," *Journal of the American Statistical Association*, 90 (432), 1357–63.
- De Finetti, Bruno (1937), "La Prévision: Ses Lois Logiques, Ses Sources Subjectives," *Annales de l'Institut Henri Poincaré*, 7, 1-68; translated as "Foresight. Its Logical Laws, Its Subjective Sources," in *Studies in Subjective Probability*, H.E. Kyburg, Jr. and H.E. Smokler, eds. Robert E. Krieger Publishing Company, 1980
- DeGroot, Morris H. (1970), *Optimal Statistical Decisions*. New York: McGraw-Hill.
- Diciccio, Thomas, Peter Hall, and Joseph P. Romano (1991), "Empirical Likelihood Is Bartlett-Correctable," *Annals of Statistics*, 19 (2), 1053–61.
- Doob, Joseph L. (1953), *Stochastic Processes*. New York: Wiley & Sons.
- Efron, Bradley (1987), "Better Bootstrap Confidence-Intervals," *Journal of the American Statistical Association*, 82 (397), 171–95.
- Evgeniou, Theodoros, Pontil Massimiliano, and Poggio Tomaso (2000), "Regularization Networks and Support Vector Machines," *Advances in Computational Mathematics*, 13, 1–50.
- Fader, Peter S., James M. Lattin, and John D. Little (1992), "Estimating Nonlinear Parameters in the Multinomial Logit Model," *Marketing Science*, 11 (4), 372–85.
- Ferguson Thomas S. (1967), *Mathematical Statistics: A Decision Theoretic Approach*. New York: Academic Press.

- (1973), “A Bayesian Analysis of Some Nonparametric Problems.” *Annals of Statistics*, 1, 209–230.
- Gelfand, Allen E., Adrian F. M. Smith, and Tai-Ming Lee (1992), “Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling,” *Journal of the American Statistical Association*, 87 (418) June, 523–32.
- Geisser, Seymour (1975) “Predictive Sample Reuse Method with Applications,” *Journal of the American Statistical Association*, 70 (350), 320–28.
- Gilbride, Timothy J. and Greg M. Allenby (2004), “A Choice Model with Conjunctive, Disjunctive, And Compensatory Screening Rules,” *Marketing Science*, 23 (3), 391–406.
- Good, I.J. and R.A. Gaskins (1980), “Density Estimation and Bump-Hunting by the Penalized Likelihood Method Exemplified by Scattering and Meteorite Data,” *Journal of the American Statistical Association*, 75 (369), 42–56.
- Helland, Inge S. (2004), “Statistical Inference Under Symmetry,” *International Statistical Review*, 72 (3), 409–422.
- Huber, Joel, Dick R. Wittink, John A. Fielder, and Richard Miller (1993), “The Effectiveness of Alternative Preference Elicitation Procedures in Predicting Choice,” *Journal of Marketing Research*, 30 (February), 104–114.
- Jeffreys, Harold (1946), “An Invariant Form for the Prior Probability in Estimation Problems,” *Proceedings of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 186 (1007), 453–461.

Jen, Lichung, Chien-Heng Chou, and Greg M. Allenby (2009), "The Importance of Modeling Temporal Dependence of Timing and Quantity in Direct Marketing," *Journal of Marketing Research*, forthcoming.

Kass, Robert E. and Larry Wasserman (1996), "The Selection of Prior Distributions by Formal Rules," *Journal of the American Statistical Association*, 91 (435), 1343–70.

Lenk, Peter (1992), "Hierarchical Bayes Forecasts of Multinomial Dirichlet Data Applied to Coupon Redemptions," *Journal of Forecasting*, 11, 603–619.

——— (1993), "A Bayesian Nonparametric Density Estimator," *Journal of Nonparametric Statistics*, 3, 53–69.

——— (1999), "Bayesian Inference of Semiparametric Regression," *Journal of the Royal Statistical Society, Series B*, 61, 863–79.

——— (2003), "Bayesian Semiparametric Density Estimation and Model Verification using a Logistic-Gaussian Process," *Journal of Computational and Graphical Statistics*, 12 (3), 548–65.

——— and Wayne DeSarbo (2000), "Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects," *Psychometrika*, 65 (1), 93–119.

———, ———, Paul Green, and Martin Young (1996), "Hierarchical Bayes Conjoint Analysis: Recovery of Partworth Heterogeneity from Reduced Experimental Designs," *Marketing Science*, 15 (2), 173–91.

——— and Ambar Rao (1990), "New Models from Old: Forecasting Product Adoption by Hierarchical Bayes Procedures," *Marketing Science*, 9, 42–53.

- Louviere, Jordan, David Hensher, and Joffre Swait (2000), *Stated Choice Methods: Applications in Marketing, Transportation and Environmental Evaluation*. Cambridge, MA: Cambridge University Press.
- Lloyd, Andrew J. (2003), "Threats to the Estimation of Benefit: Are Preference Elicitation Methods Accurate?" *Health Economics*, 12, 393–402.
- Mackenzie, John (1993), "Comparison of Contingent Preference Models," *American Journal of Agricultural Economics*, 75 (3), 593–603.
- Marshall, Pablo and Eric T. Bradlow (2002), "A Unified Approach To Conjoint Analysis Models," *Journal of the American Statistical Association*, 97, 459, 674–82.
- Meijer, Eric and Jan Rouwendal (2006), "Measuring Welfare Effects in Models with Random Coefficients," *Journal of Applied Econometrics*, 21 (2), 227–44.
- Ripley, Brian D. (1987), *Stochastic Simulation*. New York: John Wiley & Sons.
- Rossi Peter E., Robert E. McCulloch, and Greg M. Allenby (1996) "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15 (4), 321–40.
- Ryan, Mandy and Jenny Hughes (1997), "Using Conjoint Analysis to Assess Women's Preferences For Miscarriage Management," *Health Economics*, 6 (3), 261–73.
- Singh, Vishal P., Karsten T. Hansen, and Sachin Gupta (2005) "Modeling Preferences for Common Attributes in Multicategory Brand Choice," *Journal of Marketing Research*, 42 (May), 195–209.
- Seetharaman, P.B., Andrew Ainslie, and Pradeep Chintagunta (1999), "Investigating Household State Dependence Effects Across Categories," *Journal of Marketing Research*, 36 (November), 488–500.

- Sonnier, Garrett, Andrew Ainslie, and Thomas Otter (2007), "Heterogeneity Distributions of Willingness-To-Pay in Choice Models," *Quantitative Marketing and Economics*, 5 (3), 313–31.
- Stone, M (1974) "Cross-Validatory Choice and Assessment of Statistical Prediction," *Journal of the Royal Statistical Society, Series B*, 36 (2), 111–47.
- and Philip A. Dawid. (1973), "Un-Bayesian Implications of Improper Bayes Inference in Routine Statistical Problems," *Biometrika*, 59 (2), 369–75.
- Train, Kenneth E. and Terry Atherton (1995), "Rebates, Loans, and Customers Choice of Appliance Efficiency Level – Combining Stated and Revealed-Preference Data," *Energy Journal*, 16, (1), 55–69.
- Viscusi, W. Kipp, Wesley A. Magat, and Joel Huber (1991), "Pricing Environmental Health Risks: Survey Assessments of Risk-Risk and Risk-Dollar Tradeoffs for Chronic Bronchitis," *Journal of Environmental Economics and Management*, 21 (1), 32–51.
- Wahba, Grace (1983) "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society B*, 45 (1), 133–50.
- Zellner, Arnold (1971), *An Introduction to Bayesian Inference in Econometrics*. New York: John Wiley & Sons.

*FOOTNOTES*

1. Bayesian statisticians typically do not count parameters because Bayes' rules integrate over the parameter space. This paper will use parameter counts as an indicator of the size of the model space with the understanding that parameter counts do not enter into Bayesian inference in the same way they do with other inferential procedures, such as likelihood ratio tests.
2. For a scale from 1 to H the maximum standard deviation is  $0.5(H-1)$  when 1 and H both have probability 0.5 each.

Table 1: Error Covariance Estimation in Multivariate Normal Regression with a Sample Size of 10. The posterior means of the informative CNIG prior are closer to the true covariance matrix than the standard, inverse Wishart prior.

TRUE	True Error Covariance				Prior Specification for Variances					
	Y1	Y2	Y3	Y4	Mode	P90	a	b	Max	
Y1	.1	-.3	0	0	$\tau_1^2$	.05	1	1.956	.198	0
Y2	-.3	1	0	0	$\tau_2^2$	.5	5	2.737	2.369	7
Y3	0	0	10	30	$\tau_3^2$	5	15	7.526	47.630	20
Y4	0	0	30	100	$\tau_4^2$	50	125	9.785	589.263	200
	Inverse Wishart									
Mean	Y1	Y2	Y3	Y4	STD DEV	Y1	Y2	Y3	Y4	
Y1	.210	-.243	-.164	.203	Y1	.128	.196	.416	.912	
Y2	-.243	.881	.461	-.742	Y2	.196	.501	.808	1.695	
Y3	-.164	.461	4.868	8.611	Y3	.416	.808	2.824	5.435	12.52
Y4	.203	-.742	8.611	21.338	Y4	.912	1.695	5.435	7	
	CNIG Prior									
Mean	Y1	Y2	Y3	Y4	STD DEV	Y1	Y2	Y3	Y4	
Y1	.115	-.342	-.209	.051	Y1	.077	.256	.460	1.304	
Y2	-.342	1.455	.807	-.281	Y2	.256	.953	1.800	5.259	14.18
Y3	-.209	.807	8.953	14.916	Y3	.460	1.800	4.652	3	55.84
Y4	.051	-.281	14.916	91.072	Y4	1.304	5.259	14.183	5	

P90 is the 90<sup>th</sup> percentile of the Inverse Gamma distribution.

Table 2: Estimated Utilities from Simulated Data Using Default Prior. Brand 12 is omitted from the model. The omitted brand seems to be an outlier for small numbers of choice task per subject.

## Posterior Mean

Brand	True Utility	Number of Choice Tasks per Subject				
		2	4	6	10	20
1	0	.23	-.08	-.04	-.01	-.02
2	0	.17	-.10	-.08	-.13	-.09
3	0	.17	-.05	-.05	.05	.04
4	0	.12	.31	.25	.16	.15
5	0	-.02	.07	.00	-.03	.02
6	0	-.02	-.02	-.04	.01	-.05
7	0	.35	.09	.01	.01	-.02
8	0	.31	.21	.23	.22	.10
9	0	.43	.22	.24	.11	.05
10	0	.49	.05	.12	.07	.05
11	0	-.08	-.10	-.18	-.16	-.11
<b>12</b>	<b>0</b>	<b>-2.14</b>	<b>-.60</b>	<b>-.46</b>	<b>-.28</b>	<b>-.11</b>

## Posterior Standard Deviation

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	.68	.48	.44	.33	.33
2	.54	.45	.35	.27	.25
3	.85	.54	.46	.34	.31
4	.94	.49	.40	.34	.26
5	1.04	.49	.46	.40	.30
6	.67	.39	.31	.31	.24
7	.59	.44	.48	.32	.29
8	.69	.49	.43	.34	.31
9	.59	.39	.40	.34	.32
10	.48	.43	.43	.39	.31
11	.72	.40	.38	.37	.30
<b>12</b>	<b>2.78</b>	<b>1.22</b>	<b>.97</b>	<b>.77</b>	<b>.62</b>

Table 3. Market Shares Based for Simulated Data Using Default Prior. Brand 12 is omitted from the model. The true choice shares are approximately equal. The estimated choice share for brand 12 are unusually large for small numbers of choice tasks.

## Posterior Mean

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	8.00%	7.30%	7.80%	8.00%	8.10%
2	6.20%	7.10%	7.20%	7.00%	7.40%
3	8.20%	7.80%	7.70%	8.50%	8.60%
4	9.30%	10.80%	10.20%	9.50%	9.40%
5	9.10%	8.60%	8.20%	8.10%	8.40%
6	5.60%	7.40%	7.40%	8.10%	7.70%
7	8.00%	8.40%	8.30%	8.20%	8.10%
8	9.00%	9.70%	10.10%	10.00%	9.10%
9	8.00%	9.40%	10.10%	9.00%	8.70%
10	7.40%	8.10%	9.00%	8.80%	8.60%
11	5.90%	6.90%	6.60%	7.00%	7.40%
<b>12</b>	<b>15.30%</b>	<b>8.40%</b>	<b>7.40%</b>	<b>7.70%</b>	<b>8.40%</b>

## Posterior Standard Deviation

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	.08	.04	.04	.03	.03
2	.06	.04	.03	.02	.02
3	.09	.05	.04	.03	.03
4	.09	.06	.05	.03	.02
5	.07	.05	.05	.04	.03
6	.05	.03	.03	.03	.02
7	.09	.04	.05	.03	.03
8	.08	.06	.05	.04	.03
9	.07	.04	.04	.03	.03
10	.05	.04	.04	.04	.03
11	.07	.03	.03	.03	.02
<b>12</b>	<b>.33</b>	<b>.12</b>	<b>.08</b>	<b>.06</b>	<b>.05</b>

Table 4. Estimated Utilities from Simulated Data Using the Effects Prior. Brand 12 is omitted from the model. The estimated effect for brand 12 is comparable to the included effects.

Posterior Means

Brand	True Values	Number of Choice Tasks per Subject				
		2	4	6	10	20
1	0	.31	-.09	-.11	-.05	-.05
2	0	-.59	-.30	-.14	-.16	-.11
3	0	.07	-.17	-.09	.04	.02
4	0	.04	.30	.21	.16	.14
5	0	.35	-.14	-.02	-.03	.02
6	0	-.41	-.01	-.08	-.04	-.05
7	0	.25	.08	-.02	-.04	-.05
8	0	-.35	.16	.20	.20	.09
9	0	.12	.14	.16	.07	.03
10	0	.05	.04	.08	.02	.04
11	0	-.01	-.15	-.25	-.20	-.14
<b>12</b>	<b>0</b>	<b>.17</b>	<b>.11</b>	<b>.04</b>	<b>.03</b>	<b>.06</b>

Posterior Standard Deviations

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	.45	.44	.47	.31	.33
2	1.03	.53	.38	.28	.25
3	.82	.55	.45	.35	.30
4	.99	.48	.40	.33	.24
5	.58	.61	.44	.40	.29
6	.76	.39	.32	.29	.24
7	.54	.40	.51	.34	.30
8	1.12	.52	.45	.34	.31
9	.69	.46	.40	.32	.31
10	.54	.46	.42	.38	.29
11	.52	.39	.36	.36	.30
<b>12</b>	<b>.67</b>	<b>.42</b>	<b>.34</b>	<b>.3</b>	<b>.28</b>

Table 5. Market Shares Based for Simulated Data Using Effects Prior. Brand 12 is omitted from the model. The true market shares are approximately equal, and the estimated share for brand 12 is comparable to the included brands.

Posterior Means

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	8.84%	7.39%	7.60%	7.81%	8.03%
2	6.78%	6.36%	7.11%	6.94%	7.35%
3	9.13%	7.27%	7.68%	8.64%	8.54%
4	9.91%	11.03%	10.10%	9.66%	9.48%
5	8.33%	7.85%	8.19%	8.22%	8.46%
6	6.18%	7.89%	7.40%	7.88%	7.78%
7	8.56%	8.67%	8.50%	7.98%	7.95%
8	9.83%	9.83%	10.14%	10.06%	9.11%
9	9.43%	9.39%	9.60%	8.88%	8.58%
10	8.21%	8.52%	9.00%	8.60%	8.63%
11	6.10%	6.84%	6.33%	6.91%	7.30%
<b>12</b>	<b>8.71%</b>	<b>8.96%</b>	<b>8.34%</b>	<b>8.43%</b>	<b>8.78%</b>

Posterior Standard Deviation

Brand	Number of Choice Tasks per Subject				
	2	4	6	10	20
1	.06	.04	.04	.03	.03
2	.06	.04	.03	.02	.02
3	.08	.05	.04	.03	.03
4	.12	.06	.04	.03	.02
5	.08	.06	.04	.04	.03
6	.05	.04	.03	.02	.02
7	.09	.04	.05	.03	.03
8	.06	.06	.05	.04	.03
9	.07	.05	.04	.03	.03
10	.06	.05	.04	.04	.03
11	.06	.03	.03	.03	.02
<b>12</b>	<b>.08</b>	<b>.05</b>	<b>.03</b>	<b>.03</b>	<b>.03</b>

Table 6. Estimated Brand Effects for CBC Experiment Using Default Priors. In run 1 brand 9 was omitted, and in run 2 brand 4 was omitted. The estimated effects for these two brands change dramatically depending on whether they were included or excluded.

	Run #1 Brand 9 Omitted	Run #2 Brand 4 Omitted	Difference
Brand 1	1.64	1.57	
Brand 2	1.57	1.67	
Brand 3	.93	.98	
<b>Brand 4</b>	<b>-.10</b>	<b>-1.12</b>	<b>-1.02</b>
Brand 5	-.37	-.19	
Brand 6	-1.07	-1.46	
Brand 7	-.89	-.92	
Brand 8	.74	.66	
<b>Brand 9</b>	<b>-2.44</b>	<b>-1.20</b>	<b>-1.24</b>
	Average:		-1.13

Table 7. Simulation Fit Statistics of Comparing the True Partworths and Willingness-to-pay to their Tobit, Linexp, and Unconstrained Bayes Estimates. The fit statistics for the constrained and unconstrained partworths are similar, but they are much better for the constrained than unconstrained willingness-to-pay. The Tobit and Linexp results are similar. Different data were generated for the Tobit and Linexp examples.

Partworths	Correlation		RMSE		Correlation		RMSE	
	Tobit	Free	Tobit	Free	Linexp	Free	Linexp	Free
Brand 1	.611	.623	.901	.905	.660	.633	.909	.929
Brand 2	.741	.727	.701	.737	.743	.728	.745	.759
Brand 3	.695	.678	.596	.601	.722	.721	.599	.598
Quality	.888	.886	.453	.454	.917	.916	.391	.398
Price	.899	.89	.414	.441	.877	.876	.469	.474
Willingness to Pay	Correlation		RMSE		Correlation		RMSE	
	Tobit	Free	Tobit	Free	Linexp	Free	Linexp	Free
Brand 1	.723	.18	1.257	3.446	.759	.027	1.092	13.436
Brand 2	.661	.204	1.349	3.957	.737	.137	1.159	6.122
Brand 3	.598	.039	1.321	4.173	.813	.036	.804	10.655
Quality	.854	.081	.853	3.216	.859	.017	.688	10.747

Figure 1. Inverse Gamma Distributions for Error Variances. Inverse Gamma densities with the mode fixed to 10 have a “dead zone” between 0 and 3.

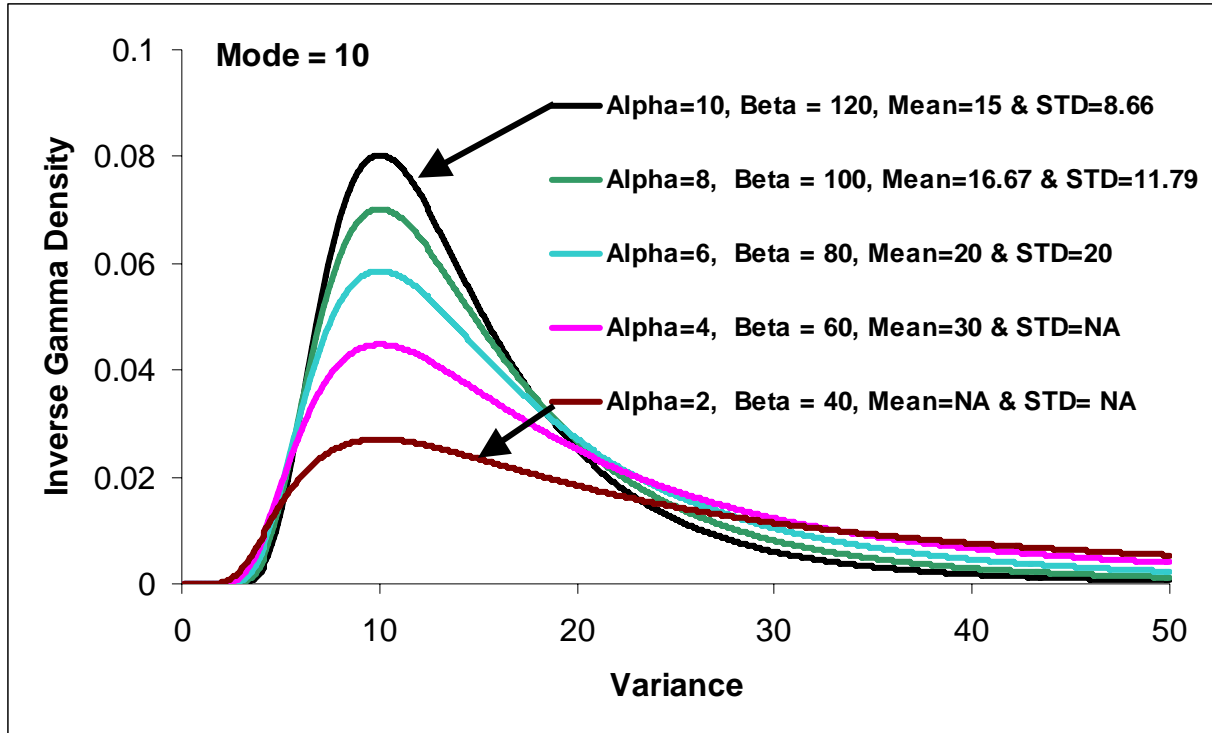


Figure 2. Differences in Posterior Means and Posterior Standard Deviations of Effects when the Effect is Omitted and Included in the Model. The effects prior is robust to the choice of the omitted effect.

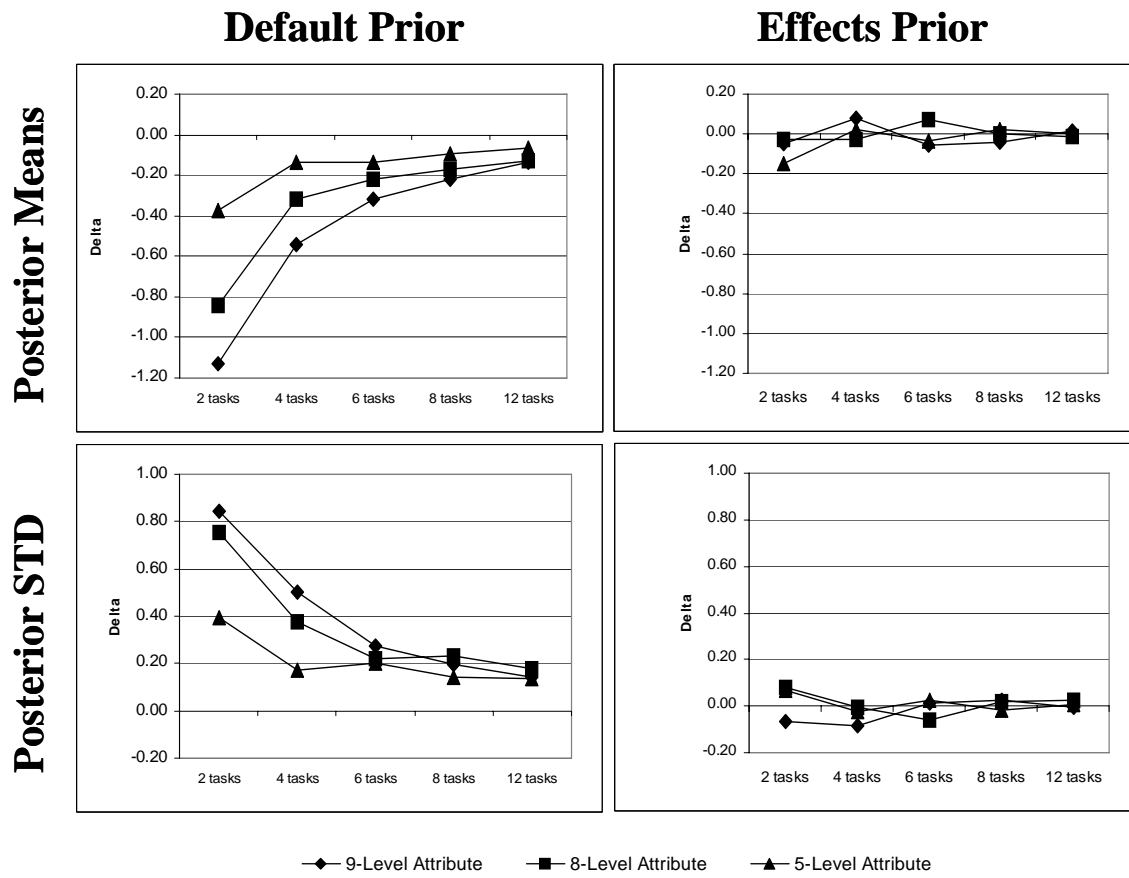
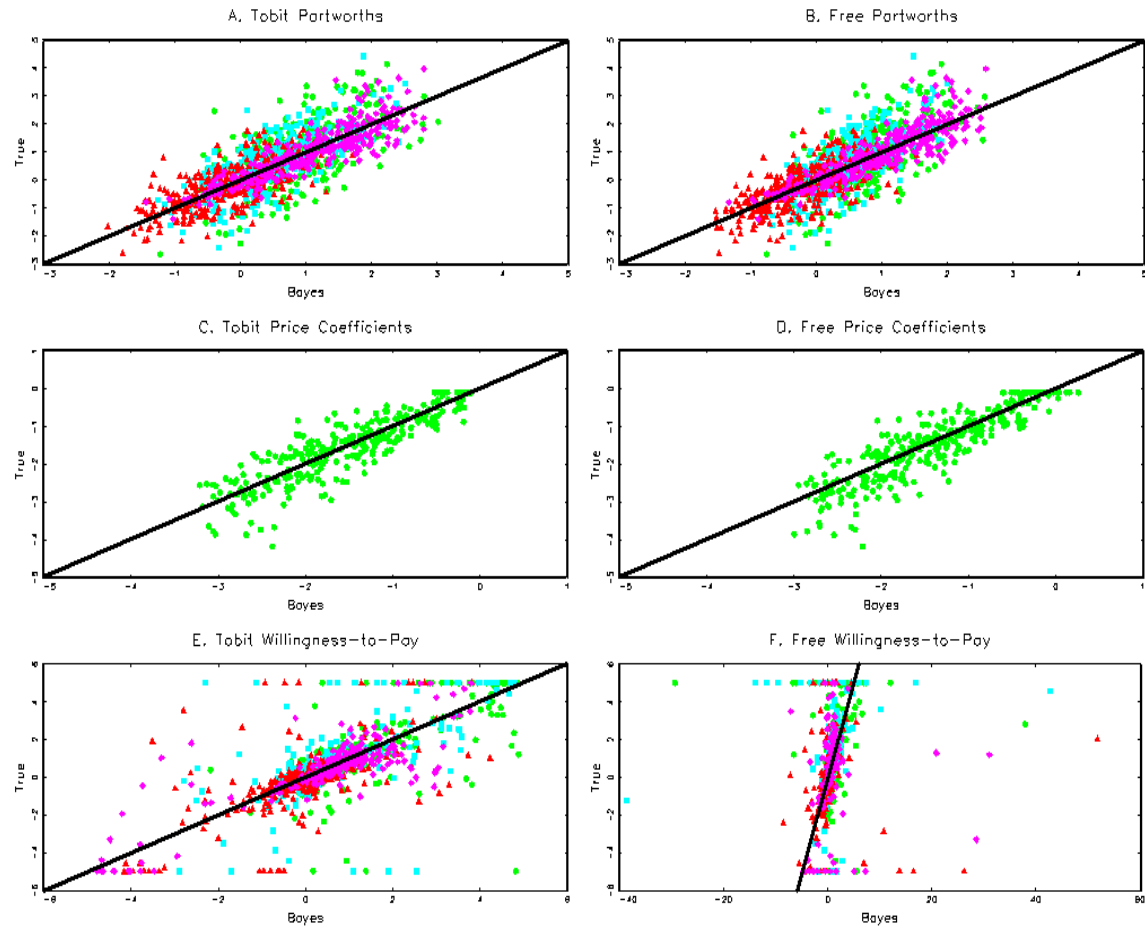


Figure 3. Limit Constraints on Partworths and Willingness-to-Pay: the True Parameters versus the Constrained and Unconstrained Bayes Estimates. The estimated partworths for the Tobit and Free models are nearly equivalent, while the estimates willingness-to-pay for Tobit model are more accurate than the Free model.



## Web Appendix

### The Value of Informative Priors in Bayesian Inference with Sparse Data

Peter Lenk and Bryan Orme

#### Web Appendix A. CNIG Prior

This appendix gives the full conditional distributions in MCMC for the CNIG prior and discusses implementation issues. The CNIG prior assumes:

$$\begin{aligned}\tau_k^2 &\sim IG\left(\frac{\alpha_k}{2}, \frac{\beta_k}{2}\right) \chi(\tau_k^2 \leq T_k^2) \text{ for } k = 1, \dots, m \\ \gamma_k &\sim N_{k-1}(\lambda_k, \Psi_k) \text{ for } k = 2, \dots, m.\end{aligned}\tag{A.1}$$

$\underline{Y}_1, \dots, \underline{Y}_n$  is a random sample for  $N_m(\mu, \Sigma)$ , and  $R_k = (Y_1 - \mu_1, \dots, Y_k - \mu_k)'$ . The full conditional for  $\tau_k^2$  given all other parameters and the data is:

$$\begin{aligned}f(\tau_k^2 | \text{Rest}) &\propto (\tau_k^2)^{-\left(\frac{\alpha_k+n}{2}+1\right)} \exp\left[-\frac{1}{2\tau_k^2}\left(\beta_k + \sum_{i=1}^n (y_{ik} - \mu_k - \gamma_k' R_{ik})^2\right)\right] \chi(\tau_k^2 \leq T_k^2) \\ \tau_k^2 | \text{Rest} &\sim IG\left(\frac{a_k}{2}, \frac{b_k}{2}\right) \chi(\tau_k^2 \leq T_k^2) \\ a_k &= \alpha_k + n \\ b_k &= \beta_k + \sum_{i=1}^n (y_{ik} - \mu_k - \gamma_k' R_{ik})^2.\end{aligned}\tag{A.2}$$

The full conditional distribution for  $\gamma_k$  given all other parameters and the data is:

$$\begin{aligned}
f(\gamma_k | \text{Rest}) &\propto \exp\left(-\frac{1}{2\tau_k^{-2}} \sum_{i=1}^n (y_{i,k} - \mu_k - R'_{i,k-1} \gamma_k)^2 - \frac{1}{2} (\gamma_k - \lambda_k)' \Psi_k^{-1} (\gamma_k - \lambda_k)\right) \\
\gamma_k | \text{Rest} &\sim N_{k-1}(\mathbf{u}_k, \mathbf{v}_k^2) \\
\mathbf{v}_k^2 &= \left( \Psi_k^{-1} + \tau_k^{-2} \sum_{i=1}^n R_{i,k-1} R'_{i,k-1} \right)^{-2} \\
\mathbf{u}_k &= \mathbf{v}_k^{-2} \left( \tau_k^{-2} \sum_{i=1}^n R_{i,k-1} (y_{i,k} - \mu_k) + \Psi_k^{-1} \lambda_k \right).
\end{aligned} \tag{A.3}$$

MCMC run times for the IW and CNIG priors are equivalent when using the Bartlett decomposition for the Wishart distribution. The Bartlett decomposition and the CNIG both generate  $m$  Gamma random variables and  $(m)(m-1)/2$  normal random variables.

The definition of  $\{\tau_k^2\}$  and  $\{\gamma_k\}$  depends on the ordering of the indices, and the CNIG prior on  $\Sigma$  is not invariant to permutation of the indices. If this is a concern, one could consider a mixture prior where each permutation of the indices has equal weight. Since the likelihood function does not depend on the permutation, one would cycle through all possible permutation in the MCMC or randomly choose one of the  $K!$  permutations on each iteration. We will not consider this extension here, though preliminary studies did not indicate a discernable inference advantage for using the permutation mixture, while MCMC run times increased.

Is it worthwhile to switch from IW to CNIG priors? The answer depends on the application. As a practical matter, CNIG distributions are easy to program, and their run times are comparable to a Bartlett's decomposition for the Wishart distribution, which is much faster than brute force methods that generate vectors of normal draws and compute their sample covariance matrix. From a theoretical perspective, CNIG is more general than the IW and can be very similar to it when one uses equal degrees-of-freedom in the CNIG. It would seem that CNIG has the edge on IW; however, in many applications the two will give similar results.

CNIG may be useful to have in the Bayesian toolbox for those situations when IW is too restrictive to encode prior beliefs.

## Web Appendix B. Effects Priors for Multi-Way Layouts with Interaction Terms

This Appendix extends the effects-coding prior covariance in the paper to more than one factor and interactions terms of all orders. In the general case, the effects-coding prior covariance is given in Equations (B.3) and (B.4) below. The basic idea is the same as the one-way layout: a symmetric prior specification on the treatment means results in symmetrical priors for included and omitted parameters. Without the appropriate notation, the general treatment is very complex and tedious; however, the results are actually very simple to implement in matrix-based programming languages.

In a multi-way layout there are  $F$  factors, and factor  $f$  has  $K_f$  levels. The treatment means are given by the combinations of the different levels of the  $F$  factors:  $\mu_{k_1, k_2, \dots, k_F}$  is the treatment mean for level  $k_1$  of factor 1, level  $k_2$  for factor 2, up to level  $k_F$  for factor  $F$ . In a full-factorial design the treatment means are arranged into a  $K = K_1 K_2 \dots K_F$  dimensional vector  $\mu$  where the last index for factor  $F$  completely cycles, before the index for Factor  $F-1$  cycles, and so on until the index for factor 1. For examples, in a two factor model with 2 and 3 levels, respectively, the treatment mean vector is  $\mu = (\mu_{11}, \mu_{12}, \mu_{13}, \mu_{21}, \mu_{22}, \mu_{23})'$ . The indexing of the treatment means, though arbitrary, needs to be consistent with the definition of main effects and interactions given below.

We identify a main effect or interaction by a subset  $E$  of the factor names. The grand mean is indicated by the null set,  $E = \{\}$ ; main effects for factor  $f$  is  $E = \{f\}$ ; a two-way interaction between factors  $f$  and  $g$  is  $E = \{f, g\}$ ; a three-way interactions among factors  $f, g,$  and  $h$  are  $E =$

{f,g,h}, and so on. Main effects and interactions are defined by averaging and mean-centering the treatment means. The averaging operator  $A_K$  and the mean-centering operator  $C_K$  are:

$$A_K = \frac{1}{K} \mathbf{1}_K' \text{ and } C_K = \left[ I_{K-1} - \frac{1}{K} J_{K-1} \quad \mid \quad -\frac{1}{K} \mathbf{1}_{K-1} \right] \text{ where} \\ \mathbf{1}_K \text{ is a } K \text{- vector of ones; } I_K \text{ is a } K \text{ by } K \text{ identity matrix, and} \\ J_K = \mathbf{1}_K \mathbf{1}_K' \text{ is a } K \text{ by } K \text{ matrix of ones.} \quad (\text{B.1})$$

$A_K$  is a 1 by  $K$  row vector, and  $C_K$  is a  $K-1$  by  $K$  matrix. For the one-way layout, the averaging operator applied to the treatment mean gives the grand mean, and the mean-centering operator gives the treatment effects. In deriving the effects-coding prior covariance matrix, we use three facts about outer products of these operators:  $A_K A_K' = 1/K$ ;  $C_K C_K' = I_{K-1} - J_{K-1}/K$ ; and  $C_K A_K' = \mathbf{0}_{K-1}$ , a  $K-1$  vector of zeros.

Mean-centering matrices  $C$  are used for factors that are included in the interaction, and averaging matrices  $A$  are used for factors that are excluded from the interaction. The transformation  $T_E$  for the treatment means to the effects parameters  $\alpha_E$  specified by  $E$  is the Kronecker product of averaging and mean-centering operators:

$$T_E = \otimes_{f=1}^F \left( A_{K_f}^{\chi(f \notin E)} C_{K_f}^{\chi(f \in E)} \right) \text{ where} \\ \chi(f \in E) = 1 \text{ if } f \text{ belongs to } E \text{ and } 0 \text{ otherwise,} \\ \chi(f \notin E) = 1 \text{ if } f \text{ does not belong to } E \text{ and } 0 \text{ otherwise,} \quad (\text{B.2}) \\ A_{K_f}^{\chi(f \notin E)} = A_{K_f} \text{ if } f \notin E \text{ and } 1 \text{ if } f \in E, \\ C_{K_f}^{\chi(f \in E)} = C_{K_f} \text{ if } f \in E \text{ and } 1 \text{ if } f \notin E.$$

Then the parameter for effect  $E$  is defined from the treatment mean as  $T_E \mu = \alpha_E$ . For instance, suppose that there are five attributes. The operator for the three-way interaction  $E = \{2, 4, 5\}$  is:

$T_E = A_{K_1} \otimes C_{K_2} \otimes A_{K_3} \otimes C_{K_4} \otimes C_{K_5}$ . The order of the averaging and mean-centering operators is important and follows the indexing convention for the components of the treatment mean vector.

A symmetrical treatment of the included and omitted effects is obtained by imposing a symmetrical prior covariance on the treatment means:  $\Sigma_\mu = \sigma_\mu^2 I_K$ . Then the covariance of the effect defined by the set  $E$  is

$$\Sigma_E = \sigma_\mu^2 T_E T_E' = \sigma_\mu^2 \left( \prod_{f \in E} \frac{1}{K_f} \right) \left( \otimes_{f \in E} \left[ I_{K_f-1} - \frac{1}{K_f} J_{K_f-1} \right] \right) \quad (\text{B.3})$$

Moreover, if the sets  $E_1$  and  $E_2$  are different effects (none identical subsets of  $1, 2, \dots, F$ ), then the covariance matrix between the two sets of effects is a matrix of zeros:  $\Sigma_{E_1, E_2} = \sigma_\mu^2 T_{E_1} T_{E_2}' = 0$ .

If a model has the effects and interactions defined by the sets  $E_1, E_2, \dots, E_g$ , then the transformation from the vector of treatment means to the full-set of model parameters is  $T = [T_{E_1}, \dots, T_{E_g}]'$  and  $T\mu = \beta$ . Consequently, the prior covariance matrix for the  $\beta$  is block diagonal:

$$\Sigma_\beta = \begin{bmatrix} \Sigma_{E_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_{E_g} \end{bmatrix}. \quad (\text{B.4})$$

Next, we present an example that illustrates using three factors:  $f$ ,  $g$ , and  $h$  where  $f$  has two levels;  $g$  has four levels, and  $h$  has three levels for a total of 24 treatments in a fully crossed experiment. Let  $\mu_{f,g,h}$  be the treatment means where  $f = 1$  or  $2$ ;  $g = 1$  to  $4$ ; and  $h = 1$  to  $3$ . The vector of treatment means is  $\mu' = (\mu_{111}, \mu_{112}, \mu_{113}, \mu_{121}, \mu_{122}, \mu_{123}, \mu_{131}, \mu_{132}, \mu_{133}, \mu_{141}, \mu_{142}, \mu_{143}, \mu_{211}, \mu_{212}, \mu_{213}, \mu_{221}, \mu_{222}, \mu_{223}, \mu_{231}, \mu_{232}, \mu_{233}, \mu_{241}, \mu_{242}, \mu_{243})$ . The grand mean is denoted by  $E = \{\}$ . The transformation from the treatment means to the grand mean is

$$T_E = A_2 \otimes A_4 \otimes A_3 = \frac{1}{24} \bar{1}_{24} \text{ so } T_E \mu = \bar{\mu}_{\cdot, \cdot, \cdot}. \quad (\text{B.5})$$

There are main effects for each factor, and we will focus on the main effects for factor g, which are defined as:

$$\alpha_{\bullet,g,\bullet} = \bar{\mu}_{\bullet,g,\bullet} - \bar{\mu}_{\bullet,\bullet,\bullet} \quad \text{where } \bar{\mu}_{\bullet,g,\bullet} = \frac{1}{6} \sum_{i=1}^2 \sum_{k=1}^3 \mu_{i,g,k} \quad \text{for } g = 1, 2, \text{ and } 3. \quad (\text{B.6})$$

Because effects sum to zero, there are 3 linearly independent main effects for factor g.

Rearranging terms, the effect for the first level is the following contrast of the treatment means:

$$\alpha_{\bullet,1,\bullet} = \frac{1}{24} (3\mu_{111} + 3\mu_{112} + 3\mu_{113} - \mu_{121} - \mu_{122} - \mu_{123} - \mu_{131} - \mu_{132} - \mu_{133} - \mu_{141} - \mu_{142} - \mu_{143} + 3\mu_{211}$$

+ 3\mu\_{212} + 3\mu\_{213} - \mu\_{221} - \mu\_{222} - \mu\_{223} - \mu\_{231} - \mu\_{232} - \mu\_{233} - \mu\_{241} - \mu\_{242} - \mu\_{243}). In the notation of this

section,  $E = \{g\}$ , and the transformation is  $T_E = A_2 \otimes C_4 \otimes A_3$  or

$$24 * T_E = \begin{pmatrix} 3, 3, 3, -1, -1, -1, -1, -1, -1, -1, -1, -1, 3, 3, 3, -1, -1, -1, -1, -1, -1, -1, -1, -1 \\ -1, -1, -1, 3, 3, 3, -1, -1, -1, -1, -1, -1, -1, -1, 3, 3, 3, -1, -1, -1, -1, -1, -1, -1, -1 \\ -1, -1, -1, -1, -1, -1, 3, 3, 3, -1, -1, -1, -1, -1, -1, -1, -1, -1, 3, 3, 3, -1, -1, -1 \end{pmatrix} \quad (\text{B.7})$$

The first row corresponds to the coefficients in the definition of  $\alpha_{\bullet,1,\bullet}$ . The other main effects

$\alpha_{\bullet,2,\bullet}$  and  $\alpha_{\bullet,3,\bullet}$  can be shown to be the other rows of  $T_E$ .

The interaction effects between factors f and h are defined by:

$$\begin{aligned} \gamma_{f,\bullet,h} &= \bar{\mu}_{f,\bullet,h} - \bar{\mu}_{f,\bullet,\bullet} - \bar{\mu}_{\bullet,\bullet,h} + \bar{\mu}_{\bullet,\bullet,\bullet} \\ \bar{\mu}_{f,\bullet,h} &= \frac{1}{4} \sum_{g=1}^4 \mu_{f,g,h}; \quad \bar{\mu}_{f,\bullet,\bullet} = \frac{1}{12} \sum_{g=1}^4 \sum_{k=1}^3 \mu_{f,g,k} \\ \bar{\mu}_{\bullet,\bullet,h} &= \frac{1}{8} \sum_{j=1}^2 \sum_{g=1}^4 \mu_{j,g,h}; \quad \bar{\mu}_{\bullet,\bullet,\bullet} = \frac{1}{24} \sum_{j=1}^2 \sum_{g=1}^4 \sum_{k=1}^3 \mu_{j,g,k} \end{aligned} \quad (\text{B.8})$$

with the constraints  $\sum_{j=1}^2 \gamma_{j,\bullet,h} = 0$  for  $h = 1$  to  $3$ ; and  $\sum_{k=1}^3 \gamma_{f,\bullet,k} = 0$  for  $f = 1$  and  $2$ . There are 6

interactions and 4 linearly independent constraints for  $2 = (2-1)(3-1)$  degrees-of-freedom. By diligently rearranging terms in the definition of the interaction, one obtains the contrasts:

$\gamma_{1,\bullet,1} = \frac{1}{24} (2\mu_{111} - \mu_{112} - \mu_{113} + 2\mu_{121} - \mu_{122} - \mu_{123} + 2\mu_{131} - \mu_{132} - \mu_{133} + 2\mu_{141} - \mu_{142} - \mu_{143} - 2\mu_{211} + \mu_{212} + \mu_{213} - 2\mu_{221} + \mu_{222} + \mu_{223} - 2\mu_{231} + \mu_{232} + \mu_{233} - 2\mu_{241} + \mu_{242} + \mu_{243})$  and  $\gamma_{1,\bullet,2} = \frac{1}{24} (-\mu_{111} + 2\mu_{112} - \mu_{113} - \mu_{121} + 2\mu_{122} - \mu_{123} - \mu_{131} + 2\mu_{132} - \mu_{133} - \mu_{141} + 2\mu_{142} - \mu_{143} + \mu_{211} - 2\mu_{212} + \mu_{213} + \mu_{221} - 2\mu_{222} + \mu_{223} + \mu_{231} - 2\mu_{232} + \mu_{233} + \mu_{241} - 2\mu_{242} + \mu_{243})$ . The other interactions are defined from these two by using the constraints. Using the above notation, these interactions are indicated by the set  $E = \{f, h\}$ . The transformation from the treatment means to the interactions is:

$$\begin{aligned} T_E &= C_2 \otimes A_4 \otimes C_3 \\ &= \frac{1}{24} [1 \quad -1] \otimes [1 \quad 1 \quad 1 \quad 1] \otimes \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix} \\ &= \frac{1}{24} \begin{bmatrix} 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 & -1 \\ -1 & 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 & -1 & 2 & -1 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 & 1 \\ 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 & 1 & -2 & 1 \end{bmatrix} \end{aligned} \quad (\text{B.9})$$

Thus, we obtain the identity  $T_E \mu = (\gamma_{1,\bullet,1}, \gamma_{1,\bullet,2})'$ .

Lastly, there are 24 three-way interactions between  $f$ ,  $g$ , and  $h$  defined by:

$$\gamma_{i,j,k} = \mu_{i,j,k} - \bar{\mu}_{i,j,\bullet} - \bar{\mu}_{i,\bullet,k} - \bar{\mu}_{\bullet,j,k} + \bar{\mu}_{i,\bullet,\bullet} + \bar{\mu}_{\bullet,j,\bullet} + \bar{\mu}_{\bullet,\bullet,k} - \bar{\mu}_{\bullet,\bullet,\bullet} \quad (\text{B.10})$$

Summing the interactions over two of the three indices gives zero, resulting in  $(2-1)(4-1)(3-1) = 6$  linearly independent interactions. The first interaction is the following contrast of the

treatment means:  $\gamma_{1,1,1} = \frac{1}{24} (6\mu_{111} - 3\mu_{112} - 3\mu_{113} - 2\mu_{121} + \mu_{122} + \mu_{123} - 2\mu_{131} + \mu_{132} + \mu_{133} - 2\mu_{141} + \mu_{142}$

$+ \mu_{143} - 6\mu_{211} + 3\mu_{212} + 3\mu_{213} + 2\mu_{221} - \mu_{222} - \mu_{223} + 2\mu_{231} - \mu_{232} - \mu_{233} + 2\mu_{241} - \mu_{242} - \mu_{243})$ . The

three way interactions are denoted by  $E = \{f, g, h\}$ , and the transformation is:

$$T_E = C_2 \otimes C_4 \otimes C_3 = \frac{1}{24} [1 \quad -1] \otimes \begin{bmatrix} 3 & -1 & -1 & -1 \\ -1 & 3 & -1 & -1 \\ -1 & -1 & 3 & -1 \end{bmatrix} \otimes \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \end{bmatrix}, \quad (\text{B.11})$$

or  $24 * T_E =$

$$\begin{pmatrix} 6, -3, -3, -2, 1, 1, -2, 1, 1, -2, 1, 1, -6, 3, 3, 2, -1, -1, 2, -1, -1, 2, -1, -1 \\ -3, 6, -3, 1, -2, 1, 1, -2, 1, 1, -2, 1, 3, -6, 3, -1, 2, -1, -1, 2, -1, -1, 2, -1 \\ -2, 1, 1, 6, -3, -3, -2, 1, 1, -2, 1, 1, 2, -1, -1, -6, 3, 3, 2, -1, -1, 2, -1, -1 \\ 1, -2, 1, -3, 6, -3, 1, -2, 1, 1, -2, 1, -1, 2, -1, 3, -6, 3, -1, 2, -1, -1, 2, -1 \\ -2, 1, 1, -2, 1, 1, 6, -3, -3, -2, 1, 1, 2, -1, -1, 2, -1, -1, -6, 3, 3, 2, -1, -1 \\ 1, -2, 1, 1, -2, 1, -3, 6, -3, 1, -2, 1, -1, 2, -1, -1, 2, -1, 3, -6, 3, -1, 2, -1 \end{pmatrix} \quad (\text{B.12})$$

The first row of the matrix is the same as the coefficients for the  $\gamma_{1,1,1}$  coefficients. The other interactions are similarly computed. Needless to say, Equation (B.3) is much easier to code than working out interactions directly.

### Web Appendix C. Dummy Variable Coding

This appendix repeats the analysis Web Appendix B for dummy variable coding. The dummy-coding prior covariance is given in Equations (C.5) and (C.6). Effects coding is often preferred to dummy variable coding because the estimated effects for different factors are independent in linear models, while they are correlated when using dummy variable coding. The distortion in the induced prior for the treatment means is substantially less for dummy variables than effects coding. The dummy variables  $d_j$  from  $j = 1, \dots, K-1$  are defined as:  $d_j = 1$  for level  $j$  and  $d_j = 0$  otherwise. In a one-way layout the treatment means are related to the dummy parameters:  $\beta_j = \mu_j - \mu_K$  for  $j = 1, \dots, K-1$ , and  $\beta_0 = \mu_K$  is the intercept or omitted effect. The inverse transformation is:  $\mu_j = \beta_0 + \beta_j$  for  $j = 1, \dots, K-1$ , and  $\mu_K = \beta_0$ . If one uses the default prior covariance  $\Sigma_\beta = v^2 I_K$  for  $\beta = (\beta_0, \dots, \beta_{K-1})'$ , then the implied prior covariance for the treatment means is asymmetric:

$$\text{cov}(\mu) = \Sigma_\mu = v^2 \begin{bmatrix} 2 & 1 & \cdots & 1 & 1 \\ 1 & 2 & \cdots & 1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 2 & 1 \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}. \quad (\text{C.1})$$

The variance of the last treatment is smaller than the first  $K-1$  treatments. If one starts with the symmetrical covariance  $\Sigma_\mu = \sigma_\mu^2 I_K$  for the treatment means, then the implied covariance for  $\beta$  is:

$$\Sigma_\beta = \sigma_\mu^2 \begin{bmatrix} 1 & -1 & -1 & \cdots & -1 \\ -1 & 2 & 1 & \cdots & 1 \\ -1 & 1 & 2 & \cdots & 1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -1 & 1 & 1 & \cdots & 2 \end{bmatrix}. \quad (\text{C.2})$$

In choice models, it is common to set the intercept  $\beta_0$  to zero, which is equivalent to setting  $\mu_K$  to zero, in order to identify the model. However, other models do not impose this constraint, and the default prior on the model's parameters induces an asymmetrical prior on the treatment means.

The general case of  $F$  factors mimics that of effects coding with different operators:

$$e_K = \left[ \begin{array}{c|c} \mathbf{0}'_{K-1} & 1 \end{array} \right] \text{ and } D_K = \left[ \begin{array}{c|c} I_{K-1} & -\mathbf{1}_{K-1} \end{array} \right] \quad (\text{C.3})$$

Operator  $e_K$ , a row vector, picks the  $K^{\text{th}}$  treatment mean as the base treatment, and  $D_K$ , a  $K-1$  by  $K$  matrix, forms contrasts between the other treatment means and the base treatments. Properties of these operators are:  $e_K e'_K = 1$ ;  $D_K D'_K = I_{K-1} + J_{K-1}$ ; and  $D_K e'_K = -\mathbf{1}_{K-1}$ . The analysis for dummy variable coding mimics that for effects coding with  $A_K$  and  $C_K$  replaced with  $e_K$  and  $D_K$ . A major difference is that the former are orthogonal and the latter are not, which is one reason that effects coding is often used by discerning statisticians instead of dummy coding.

The operator from the treatment means to the dummy parameters for effects associated with  $E$  is:

$$T_E = \otimes_{f=1}^F \left( e_{K_f}^{\chi(f \notin E)} D_{K_f}^{\chi(f \in E)} \right). \quad (\text{C.4})$$

Assuming that the covariance of the treatment means is proportional to the identity matrix and using the properties of  $e_K$  and  $D_K$ , the covariance for the parameter  $\beta_E$  is

$$\Sigma_{\beta_E} = \sigma_{\mu}^2 T_E T'_E = \sigma_{\mu}^2 \otimes_{f=1}^F \left( I_{K_f-1} + J_{K_f-1} \right)^{\chi(f \in E)}. \quad (\text{C.5})$$

If the model has  $g$  effects, that are determined by  $E_1, \dots, E_g$ , the definition of the dummy variable coding is:  $T = [T_{E_1}, \dots, T_{E_g}]'$  and  $\beta = T\mu$ . If the covariance of the treatment means is proportional to the identity matrix, then the covariance of  $\beta$  is:

$$\Sigma_{\beta} = \sigma_{\mu}^2 T T' = \sigma_{\mu}^2 \begin{bmatrix} T_{E_1} T'_{E_1} & \cdots & T_{E_1} T'_{E_g} \\ \vdots & \ddots & \vdots \\ T_{E_g} T'_{E_1} & \cdots & T_{E_g} T'_{E_g} \end{bmatrix}. \quad (\text{C.6})$$

In contrast to the recommended prior specification for effects coding, this covariance matrix is not block diagonal.

## Web Appendix D: Linexp Transformation

The linexp transformation is an alternative to the Tobit for bounding the parameters. Similar to the Tobit transformation, it is the identity function over a wide range of the parameter space. Close to the boundaries, it is an exponential function to enforce the restrictions. Instead of mapping values outside the allowable range to the boundary value as in the Tobit, the linexp transformation gradually asymptotes to the boundary. We also compare them to better-known transformations, the exponential and logistic, for limiting the support of the parameters. The exponential and logistic transformation can greatly distort the implied distribution of heterogeneity, while the Tobit and linexp preserve it.

The positive, linexp function is:

$$T_{PLX}(x) = \begin{cases} c + (b - c) \exp\left[\frac{x - b}{b - c}\right] & \text{for } x < b \\ x & \text{for } x \geq b \end{cases} \quad (\text{D.1})$$

where  $0 < c < b$ . The linexp function is bounded below by  $c$ . It is exponential between minus infinity and  $b$ , and linear after  $b$ . The linexp function is continuous, and its left and right-hand derivatives match at  $b$ . These two properties avoid a kink at the knot  $b$ . The negative, linexp function is given in Equation (14) of the paper. The double linexp function symmetrically bounds the range of the partworts from above and below at  $\pm c$ :

$$T_{DLX}(x) = \begin{cases} -c + (c - b) \exp\left[\frac{x + b}{c - b}\right] & \text{for } x < -b \\ x & \text{for } -b \leq x \leq b \\ c - (c - b) \exp\left[-\frac{x - b}{c - b}\right] & \text{for } b < x \end{cases} \quad (\text{D.2})$$

where  $0 < b < c$ . It is exponential between minus infinity and  $-b$ , the identity function between  $-b$  and  $b$ , and negative exponential between  $b$  and infinity. The double linexp is constructed to be

continuous such that its left and right-hand derivatives match at  $-b$  and  $b$  to avoid kinks at these knots.

We will compare the linexp to Tobit functions, exponential functions, and logistic functions. The positive Tobit for  $c > 0$  is:

$$T_{PT}(x) = \begin{cases} c & x < c \\ x & x \geq c \end{cases}. \quad (\text{D.3})$$

The negative Tobit for  $c < 0$  is:

$$T_{NT}(x) = \begin{cases} x & \text{for } x \leq c \\ c & \text{for } x > c \end{cases}. \quad (\text{D.4})$$

The double Tobit for  $c > 0$  is:

$$T_{DT}(x) = \begin{cases} -c & \text{for } x < -c \\ x & \text{for } -c \leq x \leq c \\ c & \text{for } x > c \end{cases}. \quad (\text{D.5})$$

Exponential transformations are frequently used to enforce positivity constraints. We consider positive and negative shifted exponential functions to enforce lower and upper bounds and the logistic function to enforce interval bounds. The positive shifted exponential function for  $c > 0$  is:

$$T_{PX}(x) = c + \exp(x) \text{ for } -\infty < x < \infty. \quad (\text{D.6})$$

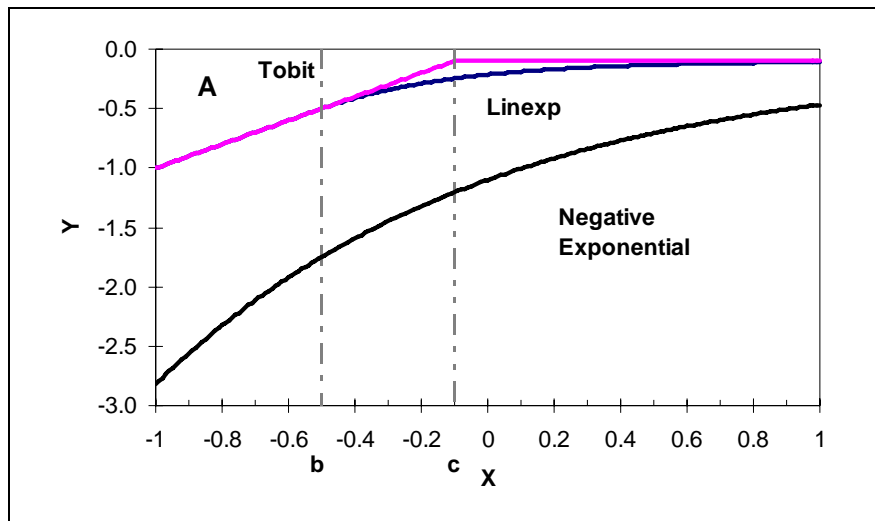
The negative shifted exponential function for  $c < 0$  is:

$$T_{NX}(x) = c - \exp(-x) \text{ for } -\infty < x < \infty. \quad (\text{D.7})$$

The logistic function or double exponential function for  $c > 0$  is:

$$T_{LGT}(x) = c \frac{\exp(x) - 1}{\exp(x) + 1} \text{ for } -\infty < x < \infty. \quad (\text{D.8})$$

Figure D.1 plots these transformations. In panel A the negative functions are bounded above by  $c=-.1$ . The negative linexp also sets  $b = -.5$ . In panel B, the double functions are bounded by  $\pm 10$  ( $c=10$ ), and the double linexp also sets  $b = 7$ . The linexp transformation are linear over most of their domains like the Tobit and unlike the exponential functions. They smoothly asymptote to their boundaries, like the exponential functions and unlike the Tobit. The Tobit transformations have kinks where the derivative does not exist, and the probability of  $X$  beyond the cutoff becomes a point mass at the cutoff. The exponential and logistic transformations are smooth; however, these functions severely distort the distribution of  $X$  if its variance is large, either by compressing the distribution or stretching it out. The logistic transformation produces an “S” shaped curve that has a very steep slope around  $x = 0$ , and very flat slopes after  $x = \pm 5$ .



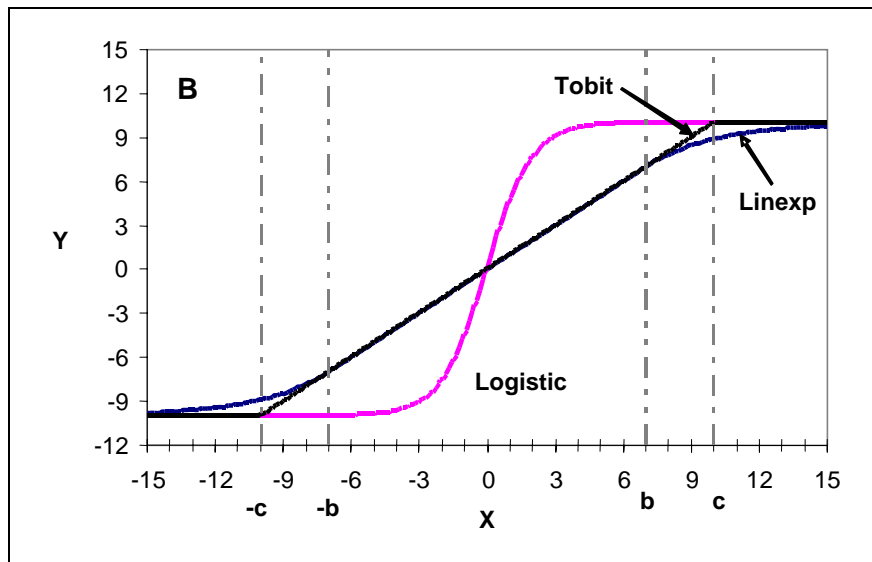


Figure D.1. Transformations to Limit Parameters. A. Negative Linexp transformation with  $b=-.5$  and  $c=-.1$ ; Tobit with  $c=-.1$ , and Negative Exponential with  $c=-.1$  B. Double Linexp transformation with  $c=10$  and  $b=7$ , Tobit with  $c=10$ , and Logistic with  $c=10$ .

Figure D.2 demonstrates the impact of these properties by displaying the histograms of the transformation in Figure D.1 to 1000 normal random deviates. Approximating histograms are superimposed. The panels in the right-hand column are the negative transformation in panel A of Figure D.1, and the panels in the left-hand column are doubly bounded functions in panel B of Figure D.2. The right-hand column applies the transformations to random draws from a normal distribution with mean  $-1$  and standard deviation  $.5$ . The left-hand column applies the transformations to random draws from a normal distribution with mean  $2$  and standard deviation  $5$ . The linexp functions in panels A and B retain the greatest fidelity to the original normal random deviates. The humps at the upper ends of the histograms are due to mapping random simulates above the upper bound  $c$  to the interval between  $b$  and  $c$ . The Tobit functions in panels C and D have spikes at the upper endpoints where the observations are truncated. Baring the

spikes, the rest of the distribution follows a normal curve. The exponential and logistic functions in panels E and F distort the normal distribution. Panel E is skewed, while panel F has a “U” shape. To be fair, the exponential and logistic transformations can be improved by introducing scaling parameters, if one knows the appropriate scale, which may be unlikely in HB models. However, they are still highly nonlinear and can result in unusual values for the partworths when applied to random normal deviates.

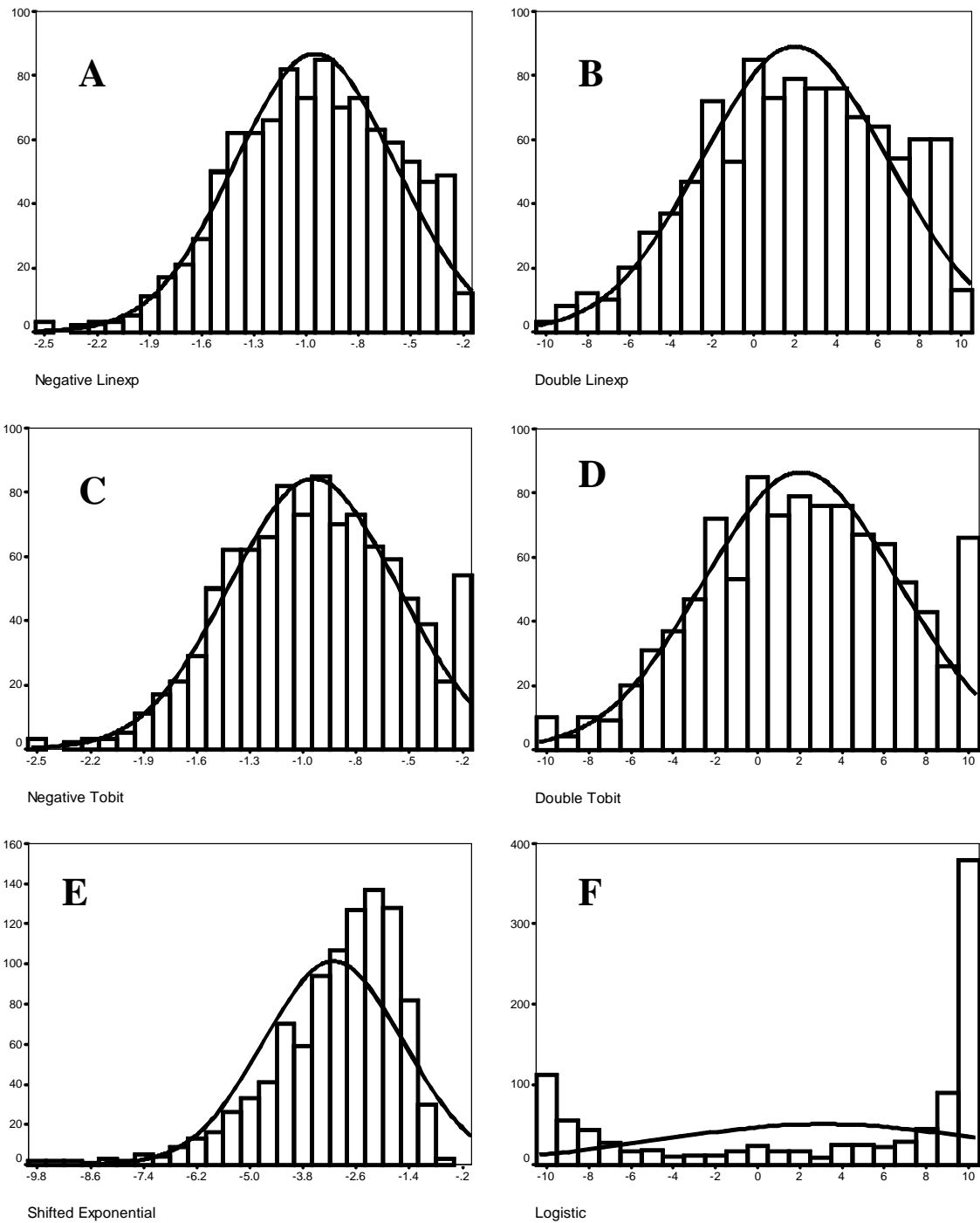


Figure D.2. Histograms of Transformed Normal Simulates. Panels A, C, and E apply the negative functions in panel A of Figure D.1 to draws from  $N(-1, .25)$ . Panels B, D, and F apply the doubly bounded functions from panel B of Figure D.1 to 1000 simulates from  $N(2, .25)$ . A and B are closest to the original normal distributions.

The lognormal distribution is frequently used to model positive parameters:  $\beta = \exp(\alpha)$  where  $\alpha$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . Rich Johnson, in a personal communication, noted that lognormal distributions tend not to work very well in CBC studies as a model of positive, parameter heterogeneity. Lognormal distributions have an unusual mathematical property that limits their utility. Their skewness parameter is proportion to their coefficient of variation (CV):

$$\text{skewness} = [\exp(\sigma^2) + 2]CV \text{ and } CV = \sqrt{\exp(\sigma^2) - 1}. \quad (\text{D.10})$$

This result implies that the distribution is “mound shaped” (small skewness) only if the CV is small. A small CV means that the standard deviation is smaller than the mean. Therefore, one obtains a mound shaped distribution only if the distribution is tight about the mean. When the variance is increased given a fixed mean, the distribution becomes more skewed. For large variances, it becomes extremely skewed. In CBC studies, this implies that there are a large number of subjects who appear as if their partworts are outliers.

The Metropolis algorithm for generating  $\{\beta_i\}$  and  $\{\alpha_i\}$  in HB multinomial logit models is a simple variation of standard code. Recall that  $\{\alpha_i\}$  has a multivariate normal distribution, and that  $\{\beta_i\}$  are the partworts in the CBC model with the relationship  $\beta_i = T(\alpha_i)$  where  $T$  is the appropriate component-wise transformation. Let  $g$  be the proposal density that is used to generate the parameters in existing software, such as a random walk. Usually  $g$  is used to generate candidates for  $\beta$ ; here it generates candidates for  $\alpha$ . Let  $L_i(\beta_i)$  be the likelihood function for the  $i^{\text{th}}$  subject, and let  $f(\alpha_i | \theta, \Lambda)$  be the heterogeneity distribution given population mean  $\theta$  and covariance matrix  $\Lambda$ . Then the Metropolis acceptance probability for moving from the current  $\alpha_i^O$  to the candidate  $\alpha_i^C$  is:

$$\psi(\alpha_i^c | \alpha_i^o) = \min \left\{ \frac{L_i[T(\alpha_i^c)] f(\alpha_i^c | \theta, \Lambda) g(\alpha_i^o | \alpha_i^c)}{L_i[T(\alpha_i^o)] f(\alpha_i^o | \theta, \Lambda) g(\alpha_i^c | \alpha_i^o)}, 1 \right\}. \quad (\text{D.11})$$

If the candidate is accepted, then set  $\beta_i^O = T(\alpha_i^C)$  and  $\alpha_i^O = \alpha_i^C$ . The acceptance probability does not include the Jacobian of the transformation because the heterogeneity distribution is for  $\alpha_i$ . Then the prior parameters  $\theta$  and  $\Lambda$  are updated in the standard way. The only coding change is substituting  $L_i[T(\alpha_i)]$  for  $L_i(\beta_i)$  in the acceptance probability.