Bayesian Applications in Marketing*

Peter E. Rossi
Anderson School of Management
UCLA

Greg M. Allenby
Fisher College of Business
Ohio State University

March 2009

Revised, May 2010

Abstract

In this chapter, we review applications of Bayesian methods to marketing problems. Key aspects of marketing applications include the discreteness of response or outcome data and relatively large numbers of cross-sectional units, each with possibly low information content. Discrete response data require the development of non-standard likelihoods and low information content requires careful use of informative priors. One particularly important form of informative prior is embodied in hierarchical models. Given the importance of the prior, it is important to assure flexibility in the prior specification. Non-standard likelihoods and flexible priors make marketing a very challenging area for Bayesian applications.

*Rossi would like to acknowledge the Kilts Center for Marketing, Booth School of Business, University of Chicago, for providing research funds. Allenby thanks the Fisher College of Business at Ohio State University for generous research support. All correspondence may be addressed to the authors at the UCLA, Anderson School of Management, 110 Westwood Plaza, Los Angeles, CA 90095; or via e-mail at perossi@gmail.com.

1 Introduction

Our approach to the application of Bayesian methods in marketing has been influenced by aspects of marketing data and the decision problems confronting both consumers and firms. While there are compelling arguments for the adoption of the Bayesian paradigm by all econometricians (see the examples and models considered by Li and Tobias (2010)), we believe that the characteristics of marketing applications make for an especially good fit with the Bayesian approach.

Marketing data originates in the decisions of individual consumers or survey respondents. Consumer data is generated by purchases of products and is often collected at the point of sale. At this level, the data is fundamentally discrete with a modal value of 0 in the sense that consumers select only a tiny subset of the available products at any one purchase occasion. Some survey methods confront consumers with a choice from among a set of products as a way of indirectly measuring consumer preferences. This data is less sparse than actual purchase data but is still discrete. Other surveys ask consumers to reflect on their use of products or exposure to media and have questions that ask consumers which of a large set of products or media to they "regularly purchase" or "regularly view." Still other survey methods collect ordinal data by using questions on a discrete rating scale. Thus, modeling of consumer level purchase or survey data requires models that allow for sparse and discrete outcome variables. Demand models that give rise to a mixture of corner and interior solutions are discussed in section 2.

Not all marketing data is available at the consumer level. Purchase data is frequently aggregated up to a store or market level. This removes some of the discreteness from the data, though even at the store level many products are sold only infrequently. However, this poses the question of what models are appropriate for aggregated data. An approach championed by some is to take the aggregation seriously by starting from a disaggregate model of sales and adding up or integrating over a distribution of consumer types. In section 2, we consider models of aggregate market shares which are derived from individual choice.

Although much data in marketing is fundamentally discrete, standard models of choice

between mutually-exclusive alternatives are not always appropriate. Consumers may not regard products, even in a narrowly defined product class, as close substitutes. For example, we often see consumers simultaneously purchasing multiple varieties of the same product. We also see consumers "trading-up" to higher quality products in response to price discounts or changes in budget allocation. Standard linear or homothetic models of choice cannot capture either phenomenon.

When consumers are confronted with a very large number of choice alternatives, they adopt decision rules to narrow the set of products under consideration. In some situations, this can be modelled formally by some sort of search process. In others, the set of products under consideration can be determined by more informal heuristics. In either case, we might adapt a more standard choice model to include a screening rule process.

Our focus on developing non-standard models for consumer purchase data is driven by the need to accommodate not just important aspects of the data itself but by the need to evaluate marketing actions. For the evaluation and improvement of marketing policies, we must postulate a model of the decision process rather than simply a descriptive statistical model for the data. Much of our research has been to explore alternative specifications of the consumer decision process, including not only the utility function over consumption of products but also the context of the decision.

In section 3, we consider more standard statistical approaches that generate discreteness by applying a censoring function to underlying continuous latent variables. This approach generates models that can be employed in situations where more descriptive models are required. In addition, the MCMC methods appropriate for models with latent variables have general usefulness for models that are derived from demand theory.

Most of marketing is predicated on the assumption that consumers differ in their evaluation of products and in their reaction to the marketing environment. Our own experience has been that this finding of large differences between consumers is pervasive. Thus, any model of purchase and consumer behavior must accommodate a rich specification for heterogeneity. Models of heterogeneity are discussed in section 4.

If the firm wants to customize its actions to specific customers or groups of customers, models of heterogeneity are doubly important. When we began our research program in marketing in the late 1980s, customization at a low level of aggregation was mostly of academic interest. However, it is now widely recognized as a potential source of value for the firm. Incorporating consumer heterogeneity requires inference for parameters at the lowest level of aggregation. This poses a challenge to standard non-Bayesian econometric methods which are focused mostly on common parameters and view individual level parameters as incidental. Another implication of the fact that we want to include parameters at the lowest level of aggregation (which we term the "unit-level") is that we are unlikely to ever have a great deal of information for the unit-level parameters. This provides a strong motivation for the use of Bayesian methods and informative priors.

As firms become increasingly sophisticated in monitoring the environment, the exogeneity assumption for marketing mix variables can be compromised. Most of the sales response or demand models are built to assume marketing mix variables are like any other covariate. It is common to condition on these variables and view them as chosen independent of the marketing environment. If a marketing mix variable (such as price) is set by the firm with partial knowledge of the unobserved portion of demand, then this assumption of independence is violated. In these situations, the conditional model of sales response must be augmented with a model of how the x variables are determined. The unobserved portion of demand can include common demand shocks, unobserved advertising or promotional activities, omitted variables capturing dynamic considerations, and unobserved product or quality characteristics. An equally important complication occurs when the firm is setting the level of the marketing mix variables with partial information about the responsiveness of the consumer to the x variables (that is when x is a function not of some portion of the error term but of the response coefficients). This later situation holds out the potential not only to bias the analysis but introduces a new source of additional information regarding the parameters of the purchase or sales model. We will explore both of these issues in section 5.

The need for non-standard models which accommodate discreteness and consumer het-

erogeneity is met by a particular set of Bayesian computational methods and models. We emphasize the need for proper, informative priors. As discussed in section 4, hierarchical models embody a form of informative prior. Bayes estimators based on informative priors display well-known shrinkage properties which ensure excellent sampling properties and numerical stability in computing. In some contexts, the use of informative priors is essential for the development of Bayesian procedures with desirable computational and convergence properties. We will see examples of this with various probit models in section 3. In these models, it can be advantageous to work in the unidentified parameter space. This is an option that is available to the Bayesian but not the frequentist. That is, we can set up computational methods that navigate the unidentified space but report on the distribution of identified parameters only. This approach requires an proper prior, however, and illustrates yet another advantage that is available with informative priors. From a purely computational point of view, proper priors modify likelihoods with singularities or near singularities, producing well-behaved posteriors. We do not require numerical linear algebra methods which are robust to ill-conditioned matrices as all posterior computations are done with properly conditioned matrices.

All of the methods and models discussed in this chapter are implemented in our contributed R (R (2009)) package, *bayesm* (Rossi and McCulloch (2008)). As required, we will cite the appropriate functions that implement a specific model. To obtain the package, install R (search for "CRAN" on the web) and use the "install packages" menu item to install, *bayesm.*

2 Demand Models

In this section, we will outline several utility specifications that incorporate discreteness and other important aspects of consumer decisions. We start with a specification of the direct utility function over product consumption and a set of random utility error terms. We prefer to derive models that are consistent with these primitives rather than simply specifying an ad-hoc statistical model.

When faced with discrete (but not necessarily multinomial data), many researchers at-

tempt to adapt a standard multinomial logit model (MNL) to the problem. Covariates such as price, product characteristics, and demographic variables are added as explanatory variables in the MNL model without a discussion of how these variables affect the consumer decision problem. There is some justification in adding price through an appeal to an indirect utility specification, but it is not clear as to how to enter product characteristics. Certainly, product characteristics may ultimately affect demand, but it is an open question as to how this occurs. Demand might be derived from preferences defined over characteristics or product characteristics could influence the decision process. For example, a product characteristic might be used as the basis of a screening rule to determine which products are entered into consideration for purchase (Gilbride and Allenby (2004)). If utility is defined over product characteristics only, this will give rise to a very different model of demand than a standard logit model[1].

For example, consider the situation in which some consumers purchase multiple varieties of the same product, such as a soft drink. The standard MNL is a model of mutually exclusive choices in which products are assumed to be highly substitutable. Clearly, purchase of soft drink varieties does not conform to the assumptions made by the MNL model. Some researchers might simply delete "troublesome" consumers or purchase occasions on which multiple products are purchased. Others might redefine the products by aggregation so as to remove the problem; for example, one might aggregate all of the colas into one product and all of the lemon-lime drinks into another. Still others, might simply ignore the problem and fit the MNL model by a non-likelihood based method such as the Generalized Method of Moments (GMM). These researchers never realize they are applying a model with zero likelihood to their data.

On the other hand, our approach is to think carefully about the decision process and utility formulation. Demand for variety might stem from multiple consumption occasions (Dubé (2004)) or simply from a non-standard utility function which allows varieties of soft drinks be imperfect substitutes. Such a utility function can give rise to multiple varieties

[1]For an extensive discussion of formulating economically consistent models of demand see Chandukala, Kim, Otter, Rossi, and Allenby (2007).

being purchased with others at zero demand (Kim, Allenby, and Rossi (2002)); that is the vector of demand in the product category can have zero entries and more than one non-zero value.

We start with a specification of the generic demand problem. In most marketing applications, attention is restricted to a subset of goods. Usually this set of products is considered to have inter-related demand and are sometimes called a product category. Here we will refer to the set of products as a demand group. In most cases, the products can be thought of as reasonably close substitutes but we might also consider the possibility there are complementarities. We recognize that consumers are deciding how much to consume or purchase not only from this set of products but also for other products outside this group for which we have little or no data or are unwilling to model explicitly. We can think of the set of products not explicitly modeled as a composite good or outside alternative and view the consumers as choosing both among the products in the category and also between this product group and the outside good.

For most marketing applications, it is not useful to take the outside alternative as literally all other products and services or even all other products within a store or shopping outlet. This would imply that the reason a consumer is not observed to purchase from the demand group at a given occasion is that the utility that can be attained from purchase in the demand group is less than that which can be obtained from the outside alternative. In most applications, a narrower definition of the outside good is a appropriate. For example, we might consider a subset of soft drinks for demand modeling. The outside good could be soft drink products outside this set or, more generally, beverages that might be regarded as potential substitutes. Defining the outside good as all of a very broad class of products puts a burden on the model to capture patterns of purchase incidence that are motivated by factors outside the model. In many applications in marketing, the analysis conditions on expenditures in the product category and there is no outside good. The limitation of the conditional approach is that the effect of changes made in the product set or marketing mix on category demand cannot be measured.

If x denotes a vector of consumption amounts for K products and z represents the outside alternative, we can formulate the consumer problem as:

$$\max_{x,z} U(x, z|\theta_i) \quad subject\ to\ p'x + z \leq E \tag{2.1}$$

Here we normalize so that the price of the outside good is 1 or that all prices are relative to the outside good. E denotes the expenditure allocation for the demand group. The utility function, $U()$, is indexed by possibly household specific parameters, θ_i. The model in (2.1) does not provide an explicit role for any marketing mix variable other than price. In particular, product attributes are not explicitly incorporated into the model. For many consumer products, it is difficult to measure more than a small subset of product attributes. Again, the soft drink example serves us well. We could define a cola attribute but we don't expect to be able to differentiate between Pepsi and Coke using product attributes. In many instances, it may be useful to think of each product as having a unique set of attributes. Other researchers partition the attributes into those which are observed by the econometrician and those that are unobservable (see, for example, Berry, Levinsohn, and Pakes (1995)). Conjoint analyses (Orme (2009)) define product explicitly in terms of attributes and seek to direct measures of utility over attributes. The influence of promotional and advertising variables requires additional assumptions about how these variables are incorporated into consumer decisions. A simple model of advertising would be that advertising exposure enhances the marginal utility of the product.

The problem in (2.1) is deterministic if the U function is known. Clearly, decisions by consumers are not perfectly predictable given only prices. For this reason, unobservables or errors are introduced into the model. The standard random utility interpretation assumes that these errors represent an unobservable utility component - consumers make deterministic demand decisions but that this unobservable to the econometrician. The interpretation of these unobservables depends in part upon their specification. We could introduce errors into utility as follows:

$$U_{i,t}^j = \overline{U}_{i,t}^j e^{\epsilon_{ijt}} \tag{2.2}$$

$U^j_{i,t}$ is the marginal utility associated with good i for consumer j on purchase occasion t. $\bar{U}$ is the deterministic part of utility function and typically would be parameterized by (possibly) consumer-specific parameters, θ_j. $\{\epsilon_{1jt}, \ldots, \epsilon_{Kjt}\}$ represents the unobservable component of utility for each of the K products in the demand group at a specific purchase occasion. The functional form of (2.2) insures that marginal utility is always positive. These errors reconcile the observed demand with that predicted given knowledge of the utility parameters and prices and form the basis for the likelihood function or the observed joint distribution of demand.

Consumer heterogeneity can be viewed as an error component that is constant across purchase occasions for the same consumer. This would create what appears to be correlation between observations for the same consumer. Our approach is different. We will assume that, conditional on consumer-specific utility parameters, θ_j, the error terms are independent across consumers. From this point on, we will suppress the individual j subscripts. We will return to the modeling of consumer differences in section 4.

The likelihood for the model is derived by considering the mapping from the error terms to the quantity demanded. If we substitute the marginal utility errors into the demand problem in (2.1), then demand, $y^*_t = (x^*_t, z^*_t)$, is a function of the marginal utility given the vector of prices and E.

$$y^*_t = f\left(\epsilon_t | \theta, p_t, E\right) \tag{2.3}$$

If the utility function is of a form which allows for corner solutions, then the quantity demanded can have a mixture of discrete and continuous distributions. The point masses in the quantity demanded must be computed by integrating over the appropriate region of the error term space which is consistent with a particular configuration of zero and non-zero demand. For example, if we observed a positive quantity of one of the brands purchased, the likelihood would receive a contribution from the density of continuous demand as well as a point mass for the probability of non-zero demand for this good. Evaluation of the likelihood would involve evaluation of the Jacobian of the transformation from ϵ to y^* as well as the computation the probabilities of various sets in the error space (see discussion in section on the demand for variety below).

Specification of the model involves choices for the functional form of U and assumptions regarding the joint distribution of the error terms. The simplest possible assumption is that the error terms are independent across consumers, purchase occasions and products (see, for example, Guadagni and Little, 1983). There are good reasons to doubt whether the assumption of independence is appropriate. For example, if there exists an unobservable product attribute, then there is a component of the error term that varies across products but is common across consumers and time periods. This induces correlation of errors across products. Aggregate demand shocks might be common across all consumers but vary from period to period, again creating a source of dependence.

Choices of the utility functional form reflect important aspects of the marketing problem. A useful simplification is to write the utility function as nesting a sub-utility function for the products in the demand group and a bivariate utility over the sub-utility function and the outside good.

$$U(x, z) = V(u(x), z) \tag{2.4}$$

This is known as a separable utility function. Under these conditions, the utility maximization problem breaks into two parts. A decision is made as to how to allocate expenditure between the outside good and the demand group. If a non-zero allocation of expenditure is made to the demand group, then a further decision is made as to how to allocate this expenditure among the products.

2.1 Linear Utility

One simple case of the separable utility function in (2.4) is a linear utility.

$$U(x, z) = \psi' x + \psi_z z \tag{2.5}$$

The linear utility assumption means that only one good (including the outside good) will be purchased as there can be no tangencies of the budget set with the indifference curves. Thus, this specification produces a mutually exclusive choice model. We should never see both the

outside good and one or more products from the demand group being chosen. However, in many applications this is ignored and the observations with a purchase from the demand group and the outside good are coded as only a purchase from the demand group.

Demand models based on the linear utility in (2.5) have an especially simple role for the expenditure variable. As long as there is sufficient budget for the purchase of the goods, the quantity demanded will be E/p and the product chosen will be the product with the highest ratio of marginal utility to price, ψ_j/p_j. As E increases, the product chosen will not change (see Allenby and Rossi, 1991). This property is illustrated in Figure 1. Linear utility is an example of a homothetic utility function. A homothetic utility function is a monotone transformation of a function that is homogenous of degree one. This means that the slope of the indifference curves is constant along any ray through the origin. As shown in (Allenby, Garratt, and Rossi (2010)), the assumption of homotheticity implies that ratio of demand for any two products is independent of E.

The likelihood for demand derived from the linear utility model is derived via a standard random utility argument as discussed above for the general case. We assume that marginal utility has a unobservable or "random" component. Assumptions regarding this component yield various possible likelihood functions. For the linear utility model, the likelihood is simply the probability of purchase of observed product choice. Quantity demanded does not provide any more information unless E is regarded as an unknown parameter. The probability of choice is derived from the distribution of the error terms which represent the unobservable component of utility. For expositional simplicity, we will fold the outside good into argument of the utility function and denote this vector as, y. $U(y) = \psi' y$. If we introduce the errors into marginal utility as in $\psi_{jt} = \psi_j e^{\epsilon_{jt}}$, then the first order conditions for choice of brand i are given by:

$$
\begin{aligned}
Pr(i) &= Pr[y_i^* > 0] && (2.6)\\
&= Pr\left(\frac{\psi_{i,t}}{p_{it}} > \frac{\psi_{j,t}}{p_{j,t}}\right) = Pr\left(ln\psi_i - lnp_{it} + \epsilon_{it} > ln\psi_j - lnp_{jt} + \epsilon_{ij}\right)\ \forall j \neq i && (2.7)\\
&= Pr\left(\epsilon_{jt} < V_{it} - V_{jt} + \epsilon_{it}\right)\ \forall j \neq i && (2.8)
\end{aligned}
$$

Here $V_{jt} = ln\psi_j - lnp_{jt}$. y_i^* denotes the quantity demanded. In order to compute the probabilities defined by the inequalities in (2.6), some assumptions have to be made regarding the distribution of the errors. If we assume that the errors are independent and identically distributed over each of goods and independent across time, then the probabilities can be expressed in terms of integrals of the cumulative distribution functions of the errors (suppressing the t subscript).

$$
\begin{aligned}
Pr\left(i\right) &= Pr(y_i^* > 0) && (2.9)\\
&= \int_{-\infty}^{+\infty} \left(\prod_{j \neq i} \int_{-\infty}^{V_i - V_j + \epsilon_i} \pi\left(\epsilon_j\right) d\epsilon_j \right) \pi\left(\epsilon_i\right) d\epsilon_i && (2.10)\\
&= \int_{-\infty}^{+\infty} \prod_{j \neq i} F\left(V_i - V_j + \epsilon_i\right) \pi\left(\epsilon_i\right) d\epsilon_i && (2.11)
\end{aligned}
$$

$F\left(\cdot\right)$ is the CDF of the error distribution and $\pi\left(\cdot\right)$ is the density of the error terms.

The error terms are introduced to accommodate the fact that consumers do not make perfectly predictable demand decisions. The error terms can also be interpreted as a model of horizontal differentiation between products. If we consider only the deterministic part of the utility function, then the linear form implies that products are perfect substitutes. The assumption of perfect substitutability between products mean that consumers are always willing to exchange two goods at a constant rate given by the ratio of the linear coefficients. Some interpret this as that perfectly substituable products have identical sources of utility but merely differ in the units of utility obtained from each product. A simple example would different package sizes of the same product. The error term introduces a source of differentiation between the products so that consumers do not view them as identical up to a scale factor. The error terms create a demand for a product even if all consumers agree that it is dominated in terms of expected value (expected marginal utility divided by price). Because the error terms have unbounded support, there is always a non-zero probability that the consumer will purchase the product. Some criticize this assumption as unrealistic. For models involving product positioning and new products, this is particularly worrisome as

it implies that all new products, however redundant with existing products, will enhance consumer welfare. These arguments tend to favor the notion that all products are valued in some sort of characteristic space which is finite but which may not be observable. This can motivate interest in models with non iid error terms.

If the errors have an extreme value type I distribution with scale parameter σ, then the CDF given by $F(t) = \exp\left(-\frac{1}{\sigma}\exp\left(-\frac{t}{\sigma}\right)\right)$ and (2.9) has a simple closed form expression (McFadden, 1981).

$$Pr(i) = \frac{\exp\left(\frac{ln\psi_i - lnp_i}{\sigma}\right)}{\sum_{j=1}^{K+1}\exp\left(\frac{ln\psi_j - lnpj_i}{\sigma}\right)} = \frac{\exp\left(\beta_{0,i} + \beta_p lnp_i\right)}{\sum_{j=1}^{K+1}\exp\left(\beta_{0,j} + \beta_p lnp_j\right)} \tag{2.12}$$

We note that here the price coefficient is the reciprocal of the extreme value error scale parameter. In many applications of MNL models, the price coefficient is used to convert the other coefficients into a monetary value for different values of the associated explanatory variables. Sonnier, Ainslie, and Otter, 2007 demonstrate the importance for careful consideration of the prior in these computations. In particular, it may be more reasonable to assess prior on the ratio of a product attribute to the price coefficient rather than separately on the price coefficient and other coefficients.

In (2.12), there are redundant parameters as we can simply normalize with respect to any of the goods. If we normalize with respect to the $K+1$st (outside good), we obtain an expression whose parameters are identified. We have used the fact that we set the price of the outside good to 1 to derive (2.13).

$$\begin{aligned} Pr(i) &= \frac{\exp\left(\beta_{0,i} + \beta_p lnp_i\right)}{\sum_{j=1}^{K+1}\exp\left(\beta_{0,j} + \beta_p lnp_j\right)} \times \frac{\exp\left(-\beta_{0,kK+1} - \beta_p lnp_{K+1}\right)}{\exp\left(-\beta_{0,kK+1} - \beta_p lnp_{K+1}\right)} \\ &= \frac{\exp\left(\tilde{\beta}_{0,i} + \tilde{\beta}_p lnp_i\right)}{1 + \sum_{j=1}^{K}\exp\left(\tilde{\beta}_{0,j} + \tilde{\beta}_p lnp_j\right)} \end{aligned} \tag{2.13}$$

The MNL model is the only model which displays the Independence of Irrelevant Alternatives (IIA) property (McFadden, 1981). The ratio of choice probabilities for products i

and j depends only on variables and parameters for these two alternatives and *not* on the characteristics of any other choice alternative as the denominators in the choice probabilities will cancel out.

$$\frac{Pr\left(i\right)}{Pr\left(j\right)}=\frac{\exp\left(\tilde{\beta}_{0,i}+\tilde{\beta}_{p}lnp_{i}\right)}{\exp\left(\tilde{\beta}_{0,j}+\tilde{\beta}_{p}lnp_{j}\right)} \tag{2.14}$$

It is well known that this same IIA property imposes severe restrictions on the cross-price elasticities of demand of product i with respect to product j (see, for example, Train (2003), section 3.6).

$$\frac{\partial Pr\left(i\right)}{\partial p_{j}}\frac{p_{j}}{Pr\left(i\right)}=\frac{\partial Pr\left(i\right)}{\partial lnp_{j}}\frac{1}{Pr\left(i\right)}=-\beta_{p}Pr\left(j\right) \tag{2.15}$$

This manifestation of the IIA property is often called the "proportional draw" property. If the price of product j is reduced, then it will draw increased demand proportionate to it choice probability. Often this property is applied at the market level. If we observe a market consisting of a large number of identical customers, then, by the law of large numbers, market shares will be similar to the choice probabilities conditional on price. To return to our soft drink example, suppose we consider a market with Coke, RC Cola, and 7-Up. Coke and RC Cola are both cola drinks with a similar taste, while 7-Up is a lemon-lime soft drink. 7-Up is a strong national brand while RC Cola is a weaker brand with only pockets of strong regional demand. For this reason, we expect that 7-Up would have a larger share than RC Cola. The logit model would imply that that the cross-price elasticity between Coke and 7-Up would be larger than that between that between Coke and RC Cola even though we might expect Cola soft drinks to be more inter-dependent in demand.

Relaxing the IIA property has spawned interest in choice models derived with linear utility but with non-independent and non-extreme value errors. In section 3, we consider the model with correlated normal errors. However, the fundamental weakness of the logit model for many marketing applications is the use of a linear utility structure.

2.2 Non-Homothetic Utility for Multinomial Data

Consider the problem of formulating a demand model for mutually exclusive choices, but where the choice options differ in quality. For example, consumers may purchase only one of a variety of offerings in categories such as cars, vacations and electric razors. Choice is still characterized as a strict corner solution where just one of the alternatives has non-zero demand. As demonstrated earlier, this can only be ensured when indifference curves are linear. However, as the the budget allocated to the product category increases, we expect a higher demand for higher qualty goods. That is, consumers achieve higher utility by purchasing higher quality goods rather than simply consuming more of lower quality goods. We can define goods whose demand increases as expenditure increases as goods that are relatively superior to goods of lower quality for which demand declines as expenditure allocation increases. The lower quality goods are be termed "relatively inferior."

Allenby and Rossi (1991) and Allenby, Garratt, and Rossi (2010) propose an implicitly defined utility function with linear indifference curves but non-constant marginal utility.

$$\begin{aligned} U(x,z) &= ln\left(u(x)\right) + \tau ln(z) \\ u(x) &= \sum_{k=1}^{K} \psi_k(\bar{u})\, x_k \\ \psi_k(\bar{u}) &= \exp\left[\alpha_k - \kappa_k \bar{u}(x,z)\right] \end{aligned} \tag{2.16}$$

In this specification, marginal utility increases with attainable utility, $\bar{u}$. As consumers allocate more expenditure to the product class, attainable utility increases and relatively superior products increase their marginal utility relative to the inferior products. The utility function has valid linear indifference curves if $\kappa_k > 0$. Relatively smaller values of κ are associated with superior goods. Figure 2 shows the rotating indifference curves associated with this utility function. As the budget constraint is relaxed, the product chosen can shift as the relatively superior product becomes more highly valued given that the ratio of marginal utilities is not constant and increasing in attainable utility for the relatively superior products.

The non-homothetic utility function specified in (2.16) overcomes a number of the limi-

tations of homothetic utility functions. In many situations, we observe that consumers are willing to "trade-up" from lower priced and lower-quality products to higher priced and higher-quality products. For any homothetic utility system, product choice will not be altered by increases in E. This is an important consideration both across consumers and for the same consumer observed over time. Wealthier consumers may have greater E, but, more importantly, price reductions in a demand group can stimulate a demand for higher quality products in the group. This effect can be captured in the non-homothetic choice model discussed here but not in the standard, linear utility, logit formulation.

In (2.16) a standard bivariate utility over the product class and the outside good that insures an interior solution with an inside good purchased as long as the expenditure allocation is sufficient to allow for purchase. Given an assumption on marginal utility errors, we can derive the probability of choice of a brand by recognizing that only one brand will be purchased and, therefore, $E - p_i$, will be allocated to the outside alternative if brand i is chosen.

The log of utility associated with choice of brand i is:

$$ln\bar{u}^i = \alpha_i - \kappa_i \bar{u}^i + \tau ln\left(E - p_i\right) \tag{2.17}$$

where $\bar{u}^i$ is the root of the equation:

$$lnx + \kappa_i x - \alpha_i - \tau ln\left(E - p_i\right) = 0 \tag{2.18}$$

The root of (2.18) can be found by Newton's method. We note that Newton's method can be shown to converge to the unique root of (2.18).

We can derive the likelihood or choice probabilities for this model by the usual device of introducing errors into the marginal utilities.

$$\psi_{i,t}\left(\bar{u}\right) = \psi_i\left(\bar{u}\right)\exp\left(\epsilon_{it}\right)$$

The probability of selecting product i, given the computation of $\bar{u}^i$ for all choice alternatives,

is:

$$Pr\left(i\right)=Pr\left(\alpha_i-\kappa_i\bar{u}^i+\tau ln\left(E-p_i\right)+\epsilon_k>\alpha_j-\kappa_j\bar{u}^j+\tau ln\left(E-p_j\right)+\epsilon_j\ \ \forall j\ni p_j\leq E\right)$$

If the error terms are the standard extreme value type I, we get a MNL with a non-linear regression function and the restriction that a non-zero probability of choice obtained only for products that are in the feasible consumption set.

$$Pr\left(i\right)=\frac{\exp\left(\alpha_i-\kappa_i\bar{u}^i+\tau ln\left(E-p_i\right)\right)}{\sum_{j\ni p_j<E}\exp\left(\alpha_j-\kappa_j\bar{u}^j+\tau ln\left(E-p_j\right)\right)} \tag{2.19}$$

The non-homothetic choice model in (2.19) builds on th‘e simple extreme value and independent errors but gives rise to a much richer pattern of demand with non-trivial income effects. It should be emphasized that income effects occur not only across households but for the same household when there are price changes in the demand group. During sale periods, the feasible level of utility can rise and this can increase the marginal utility of higher quality brands, inducing consumers to trade up to these offerings.

2.3 Multiple Discreteness, Satiation and the Demand for Variety

While variants of the multinomial choice model (any model in which the outcome is a multinomial random variable like the MNL or MNP models) are immensely popular both in marketing and in the Industrial Organization literature, we frequently observe non-mutually exclusive choice behavior. For example, consumers are observed to buy more than one variety of yogurt or movies or CD's while it is still true that consumers rarely buy more than a tiny fraction of the available products. For this reason, some have called this type of demand, multiple discreteness. We will need a utility function which is capable of a mixture of interior and corner solutions. Figure 3 displays a non-linear utility function that is capable of both interior and corner solutions as the indifference curves intersect axes instead of the more typical case where the indifference curves are tangent to the axes. In modeling the demand for variety, we might also consider the possibility of satiation or declining marginal utility. A simple utility function

which is achieves these goals is a translated power utility, as proposed by Kim, Allenby, and Rossi (2002).

$$u(x) = \sum_{k=1}^{K} \psi_k (x_k + \gamma_k)^{\alpha_k} \tag{2.20}$$

In this model, the γ parameters serve to translate the utility so that the indifference curves can intersect the axes and result in corner solutions. The α parameters result in diminishing marginal utility. Marginal utility can be calculated from (2.20) as $u^i = \alpha_i \psi_i (x_i + \gamma_i)^{\alpha_i - 1}$.

We can derive the likelihood for this model from the Kuhn-Tucker (K-T) conditions (see, for example, Avriel (1976)) for the consumer problem (Pudney (1989)). We introduce a multiplicative error into marginal utility, $u^i = \bar{u}^i \exp(\epsilon_i) = \alpha_i \psi_i (x_i + \gamma_i)^{\alpha_i - 1} \exp(\epsilon_i)$, and differentiate the Lagrangian. x_j^* denotes the quantity demanded (optimal solutions which obeys the K-T conditions).

$$\bar{u}^i \exp(\epsilon_i) - \lambda p_j = 0 \text{ if } x_j^* > 0$$

$$\bar{u}^i \exp(\epsilon_i) - \lambda p_j < 0 \text{ if } x_j^* = 0$$

Dividing by price and taking the logs, we obtain.

$$V_j\left(x_j^* | p_j\right) + \epsilon_j = ln\lambda \quad \text{if } x_j^* > 0 \tag{2.21}$$

$$V_j\left(x_j^* | p_j\right) + \epsilon_j < ln\lambda \quad \text{if } x_j^* = 0 \tag{2.22}$$

where $V_j\left(x_j^* | p_j\right) = ln\left(\bar{u}^j\right) - ln(p_j)$.

Optimal demand satisfies the K-T conditions in equations (2.21) and (2.22) as well as the "adding-up" constraint that $p'x^* = E$. The "adding-up" constraint induces a singularity in the distribution of x^*. To handle this singularity, we use the standard device of differencing the first order conditions with respect to one of the goods. If one of the goods is the composite outside good, we can assume that this good is always consumed in non-zero quantity. Otherwise, we find one of the goods with positive demand. Without loss of generality assume

that this is good 1. The K-T conditions can now be written.

$$v_j = h_j\left(x^*, p\right) \quad \text{if } x_j^* > 0$$

$$v_j < h_j\left(x^*, p\right) \quad \text{if } x_j^* = 0$$

where $v_j = \epsilon_j - \epsilon_1$ and $h_j\left(x^*, p\right) = V_1 - V_j$ and $j = 2, \ldots, m$.

The likelihood for a given vector of demands can be constructed given an assumed distribution of the differenced errors, $\nu = \nu_2, \ldots, \nu_m$. If we assume that $\nu \sim N\left(0, \Omega\right)$, then the distribution of quantity demanded can be derived as a mixed discrete-continuous distribution. If we observe a corner (namely, 0 consumption of one or more goods), then a discrete lump of probability is introduced by the fact that there is an entire region (a subset of R^{m-1}) of the marginal utility errors consistent with this corner solution. For non-zero quantities demanded, then there will be a continuous density derived by change-of-variable methods. Suppose the first n goods have non-zero demand. The likelihood of quantity demanded has a continuous component for the first n goods, combined with the probability that the last $n+1, \ldots, m$ goods have zero demand.

$$\begin{aligned} &Pr\left(x_i^* > 0,\ i = 2, \ldots, n;\ x_i^* = 0,\ i = n+1, \ldots, m\right) \\ &= \int_{-\infty}^{h_m} \ldots \int_{-\infty}^{h_{n+1}} \phi\left(h_2, \ldots, h_n, \nu_{n+1}, \ldots, \nu_m | 0, \Omega\right) |J| d\nu_{n+1} \ldots d\nu_m \end{aligned} \tag{2.23}$$

where $\phi\left(\cdot\right)$ is the multivariate normal density, $h_j = h_j\left(x^*, p\right)$, and J is the Jacobian with elements given by

$$J_{ij} = \frac{\partial h_{i+1}\left(x^*, p\right)}{\partial x_{j+1}^*} \quad i, j = 1, \ldots, n-1. \tag{2.24}$$

Kim, Allenby, and Rossi (2002) explain how to evaluate the likelihood using the GHK method (see Keane, 1994 and Hajivassiliou, McFadden, and Ruud, 1996) to compute the required integrals (see Bhat, 2005 and Bhat, 2008 for the case of extreme value errors). R source code for this likelihood is available on the web site for Rossi, Allenby, and McCulloch (2005).

2.4 Models for Aggregate Shares

In many marketing research contexts, individual consumer level data is not available. Rather, data is aggregated over consumers such as store level or account or market level data. In particular, aggregate data is often summarized by market shares along with some market size variable. The modeling of market share data is important for the practice of marketing. One reasonable point of view is that the models for aggregate share data should be *consistent* with those postulated at the individual level even if individual consumer data is not available. For example, it is possible to take the standard multinomial logit model as the model governing consumer choice. In a market with a very large number of consumers, the market shares are the expected probabilities of purchase which would be derived by integrating the individual model over the distribution of heterogeneity. The problem is that with a continuum of consumers all of the choice model randomness would be averaged out and the market shares would be a deterministic function of the included choice model covariates. To overcome this problem, Berry, Levinsohn, and Pakes (1995) introduced an additional error term into consumer level utility which reflects a market wide unobservable. For their model, the utility of brand j for consumer i and time period t is given by

$$U_{ijt} = X_{jt}\theta_j^i + \eta_{jt} + \epsilon_{ijt} \tag{2.25}$$

where X_{jt} is a vector of brand attributes, θ_j^i is a $k \times 1$ vector of coefficients, η_{jt}, is an unobservable common to all consumers, and ϵ_{ijt} is the standard idiosyncratic shock (i.i.d. extreme value type I). If we normalize the utility of the outside good to zero, then market shares (denoted by s_{jt}) are obtained by integrating the multinomial logit model over a distribution of consumer parameters, $f\left(\theta^i|\delta\right)$, $\theta^i = \left[\theta_1^i, \ldots, \theta_J^i\right]$. δ is the vector of hyper-parameters which govern the distribution of heterogeneity.

$$s_{jt} = \int \frac{\exp\left(X_{jt}\theta_j^i + \eta_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(X_{kt}\theta_k^i + \eta_{kt}\right)} f\left(\theta^i|\delta\right) d\theta^i = \int s_{ijt}\left(\theta^i|X_t, \eta_t\right) f\left(\theta^i|\delta\right) d\theta^i$$

While it is not necessary to assume that consumer parameters are normally distributed, most applications assume a normal distribution. In some cases, difficulties in estimating the parameters of the mixing distribution force investigators to further restrict the covariance matrix of the normal distribution to a diagonal matrix (see Jiang, Manchanda, and Rossi (2009)). Assume that $\theta^i \sim N\left(\bar{\theta}, \Sigma\right)$, then the aggregate shares can be expressed as a function of aggregate shocks and the preference distribution parameters.

$$s_{jt} = \int \frac{\exp\left(X_{jt}\theta^i + \eta_{jt}\right)}{1 + \sum_{k=1}^{J} \exp\left(X_{kt}\theta_k^i + \eta_{kt}\right)} \phi\left(\theta^i | \theta, \Sigma\right) d\theta^i = h\left(\eta_t | X_t, \bar{\theta}, \Sigma\right) \tag{2.26}$$

η_t is the $J \times 1$ vector of common shocks.

If we make an additional distributional assumption regarding the aggregate shock, η_t, we can derive the likelihood. Given that we have already made specific assumptions regarding the form of the utility function, the distribution of the idiosyncratic choice errors, and the distribution of heterogeneity, this does not seem particularly restrictive. However, the recent literature on GMM methods for aggregate share models does emphasize the lack of distributional assumptions regarding the aggregate shock. We will assume that the aggregate shock is i.i.d. across both products and time periods and follows a normal distribution, $\eta_{jt} \sim N\left(0, \tau^2\right)$. The normal distribution assumption is not critical to the derivation of the likelihood; however, as Bayesians we must make some specific parametric assumptions. In theory, the GMM estimator should be robust to autocorrelated and heteroskedastic errors of an unknown form. Jiang, Manchanda, and Rossi (2009) propose a Bayes estimator based on a normal likelihood and document that this estimator has excellent sampling properties even in the presence of mis-specification and, in all cases considered, has better sampling properties than a GMM approach (see Chen and Yang, 2007 and Musalem, Bradlow, and Raju, 2009 for other Bayesian approaches).

The joint density of shares at "time" t (in some applications of aggregate share models, shares are observed over time for one market and in others shares are observed for a cross-section of markets. In the latter case, the "t" index would index markets) can be obtained by

using standard change of variable arguments.

$$\begin{aligned}\pi\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma,\tau^2\right) &= \phi\left(h^{-1}\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma\right)|0,\tau^2 I_J\right)J_{(\eta\to s)} \\ &= \phi\left(h^{-1}\left(s_{1t},\ldots,s_{Jt}|X,\bar{\theta},\Sigma\right)|0,\tau^2 I_J\right)\left(J_{(s\to\eta)}\right)^{-1} \quad (2.27)\end{aligned}$$

$\phi(\cdot)$ is the multivariate normal density. The Jacobian is given by

$$J_{(s\to\eta)} = \left\|\frac{\partial s_j}{\partial \eta_k}\right\| \quad (2.28)$$

$$\frac{\partial s_j}{\partial \eta_k} = \begin{cases}\int -s_{ij}\left(\theta^i\right)s_{ik}\left(\theta^i\right)\phi\left(\theta^i|\bar{\theta},\Sigma\right) & k\neq j \\ \int s_{ij}\left(\theta^i\right)\left(1-s_{ik}\left(\theta^i\right)\right)\phi\left(\theta^i|\bar{\theta},\Sigma\right) & k=j\end{cases} \quad (2.29)$$

It should be noted that, given the observed shares, the Jacobian is a function of Σ only (see Jiang, Manchanda, and Rossi (2009) for details).

To evaluate the likelihood function based on (2.27), we must compute the h^{-1} function and evaluate the Jacobian. The share inversion function can be evaluated using the iterative method of BLP (see Berry, Levinsohn, and Pakes, 1995). Both the Jacobian and the share inversion require a method for approximation of the integrals required to compute "expected share" as in (2.26). Typically, this is done by direct simulation; that is, averaging over draws from the normal distribution of consumer level parameters. It has been noted that the GMM methods can be sensitive to simulation error in the evaluation of the integral as well as errors in computing the share inversion. Since the number of integral estimates and share inversions is of the order of magnitude of the number of likelihood or GMM criterion evaluations, it would desirable, from a strictly numerical point of view, that the inference procedure exhibit little sensitivity to the number of iterations of the share inversion contraction or the number of simulation draws used in the integral estimates. Our experience is that the Bayesian methods that use stochastic search as opposed to optimization are far less sensitive to these numerical errors. For example, Jiang, Manchanda, and Rossi (2009) show that the sampling properties of Bayes estimates are virtually identical when 50 or 200 simulation draws are used in the

approximation of the share integrals; this is not true of GMM estimates.

2.5 MCMC Suggestions

In this section, we have introduced a variety of models designed to capture aspects of the discrete decision process at the consumer level. Even the simplest linear utility models give rise to a likelihood which is not amenable to conjugate analysis (compare to the standard linear models in Li and Tobias (2010)). The non-homothetic model (2.19), variety models (2.23), and aggregate share models (2.27) define likelihoods that must be evaluated by using numerical approximations to various integrals and, possibly, roots of equations that implicitly define the utility function (2.18). RW Metropolis methods (Chib (2010)) are ideal for these non-conjugate problems of modest dimension. For satisfactory performance, some tuning will be required for the Metropolis method. In particular, the random walk covariance matrix must be chosen with care. Recall that a general strategy is to propose candidates as follows

$$\theta^c = \theta + \epsilon \quad \epsilon \sim N\left(0, s^2 C\right) \tag{2.30}$$

For the linear utility, multinomial logit model, C can be chosen as any reasonable Hessian estimate and the Roberts and Rosenthal (2001) suggestion that $s = {}^{2.93}/\sqrt{dim(\theta)}$ will work well. The non-homothetic utility function can be tuned successful using either Hessian estimates from an optimizer or the covariance matrix of a shorter "tuning" run. The log-concavity of the standard MNL model and the near log-concavity of the non-homothetic model insure adequate performance of an initial optimizer. However, the variety model and the aggregate share model have far less regular likelihoods and should be tuned with shorter initial runs using a diagonal or identity RW increment co-variance matrices.

If the elements of the variance-covariance matrix of the normal distribution of individual parameters in the aggregate logit model (2.26) are included in the parameter vector used in a RW Metropolis step, then we must impose positive definiteness. We reparameterize so as to impose positive definiteness and use a standard RW metropolis on a unrestricted parameter

space. Positive definiteness can be imposed by writing Σ in terms of its Cholesky root.

$$\Sigma = U'U$$

$$U = \begin{bmatrix} e^{r_{11}} & r_{12} & \cdots & r_{1K} \\ 0 & e^{r_{22}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & r_{K-1,K} \\ 0 & \cdots & 0 & e^{r_{KK}} \end{bmatrix} \tag{2.31}$$

K is the number of coefficients in the micro-level choice model. The prior is assessed on the r vector which induces a prior on Σ. The relationship between the Cholesky root and the elements of Σ is order dependent. Jiang, Manchanda, and Rossi (2009) discuss an assessment procedure that induces a prior on the variance elements of Σ which has approximately the same diffusion for each element. It should be emphasized that, if diffuse priors are desired, it is a simple matter to assess a normal prior on r as long as different priors are used for the diagonal vs. off-diagonal Cholesky elements.

$$r_{jj} \sim N\left(0, \sigma^2_{r_{jj}}\right) \quad j = 1, \ldots, K \tag{2.32}$$

$$r_{jk} \sim N\left(0, \sigma^2_{off}\right) \quad k = 1, \ldots, K;\ j > k \tag{2.33}$$

Jiang, Manchanda, and Rossi (2009) show how to choose the parameters of the priors in (2.32) and (2.33) so as to achieve approximately the same prior dispersion for all elements in Σ.

3 Statistical Models for Discrete Data

In this section, we will discuss various models for discrete data that are motivated primarily by descriptive aspects of the outcome measure. In particular, we will consider multinomial and multivariate probit models based on a correlated normal error structure and a generalized model for count data (see, also, Li and Tobias, 2010).

3.1 Multinomial Probit

The Multinomial Logit Model (2.13) can be derived from linear utility and extreme value errors that are independent across choice alternatives. A natural generalization of this model would be to allow for correlation between the error terms. These correlated error terms would allow for more flexible substitution between alternatives as explanatory variables such as price vary. The proportional draw or IIA property of the logit model means that elasticities are driven entirely by one parameter. Correlation between errors allows for products that are positively correlated to have higher cross-elasticities or greater substitutability (see, for example, Hausman and Wise, 1978). One interpretation of the correlation is that it might arise from some sort of unobservable characteristic. Choice alternatives that have similar levels of this characteristic are more highly correlated. To allow for correlation in the utility errors, we move to an underlying normal latent regression model. Consider the situation in which we choose between p alternatives.

$$\begin{aligned} y_i &= f(z_i) \\ f(z_i) &= \sum_{j=1}^{p} jI\left(max\left(z_i\right) = z_{ij}\right) \\ z_i &= X_i\delta + v_i \quad v_i \sim N\left(0, \Omega\right) \end{aligned} \tag{3.1}$$

X_i contains explanatory variables which can be thought of as of two types: i. alternative specific information such a price or other marketing mix variables and ii. covariate or "demographic" information characterizing the respondents or consumers whose choices are being observed. In many applications, we also include an alternative specific intercept term that can be interpreted as some sort of vertical quality measure.

$$X_i = \left[\begin{array}{cc} (1, d_i) \bigotimes I_p & A_i \end{array} \right]$$

d_i is a vector of covariates (for example, demographic variables) and A_i contains observations on the alternative specific variables such product attributes.

The model in (3.1) has both a location and scale invariance identification problem. Location invariance means that we can add a scalar random variable u to z and not affect the alternatives chosen, i.e. $f(z_i + u\iota) = f(z_i)$. With a full-covariance matrix, the models are observationally equivalent, as $Var(z_i + u\iota | X_i, \delta) = \Omega + \sigma_u^2 I$ and $Var(z_i | X_i, \delta) = \Omega$, both are unrestricted covariance matrices. This identification problem can be restated as that all latent comparisons are relative and there has to be some sort of normalization. With a full-covariance structure, the most convenient way is to difference the latent system by subtracting one of the alternatives from all of the others. With a restricted covariance structure such as a diagonal Ω, it is possible to achieve identification simply by fixing one of the alternatives to have a zero intercept and fixing one of the diagonal elements of Ω.

The differenced system is written

$$w_i = X_i^d \beta + \epsilon_i \quad \epsilon_i \sim N(0, \Sigma) \tag{3.2}$$

where

$$w_{ij} = z_{ij} - z_{ip}, \quad X_i^d = \begin{bmatrix} x'_{ij} - x'_{ip} \\ \vdots \\ x'_{i,p-1} - x'_{ip} \end{bmatrix}, \quad \epsilon_{ij} = v_{ij} - v_{ip} \tag{3.3}$$

The differenced system is now $p-1$ dimensional; $y_i = k, k = 1, \ldots, p-1$ if $max(w_i) = w_{ik}$ and $y_i = p$ if $w_i < 0$. The ***bayesm*** function, `createX`, can be used to automatically configure X matrices using information on alternative specific and non-alternative specific explanatory variables, include intercepts, set the base alternative, and difference.

Even the differenced system in (3.2) is not identified as the system still exhibits scale invariance. If we scale the vector of differenced latents, w_i, by multiplying by a positive constant, then we still leave the observed choice unchanged. In the classical literature, this is typically handled by normalizing an element of Σ to be one. The identified parameters in (3.2) are $\tilde{\beta} = \beta / \sqrt{\sigma_{ii}}$ and $\tilde{\Sigma} = \Sigma / \sigma_{11}$.

However, in the Bayesian approach, it is not necessary to impose identification restric-

tions prior to the analysis of the posterior. One approach is to put informative priors on the unidentified parameter space and recognize that certain functions of the parameters have a posterior that is influenced by both the prior and the likelihood and other functions are only influenced by the prior. There are two advantages of this approach: i. standard normal/Inverted Wishart priors can be used for the unidentified parameters, ii. the MCMC method that navigates the unidentified space has superior mixing properties as discussed in Rossi, Allenby, and McCulloch (2005), section 4.2. The standard priors are given by

$$\beta \sim N\left(\underline{\overline{\beta}}, \underline{A}^{-1}\right) \quad \Sigma \sim IW\left(\underline{V}, \underline{\nu}\right) \tag{3.4}$$

The priors in (3.4) are on the unidentified parameter space. These can either be regarded as a legitimate statement of prior beliefs or as a device to induce a prior on the space of identifed parameters $\widetilde{\beta}, \widetilde{\Sigma}$. The user of the MCMC algorithm with priors on the non-identified parameters must check this induced prior to see that it is reasonable. In many applications, all that is desired is a relatively diffuse but proper prior. It is a simple matter to check the induced prior by simulating from (3.4) and transforming to the identified parameters. Marginals of the prior on the identified prior can be inspected to see if they represent the investigator's beliefs. In figure 4, we plot the prior distribution for the identified parameters in the MNP model using the defaults for the `rmnpGibbs` procedure in ***bayesm***. The top panel displays the prior induced on an identified regression coefficient, the middle panel shows a representative covariance element of the identified covariance matrix $\tilde{\Sigma}$, and the bottom panel shows the prior for a representative correlation. It is a simple matter to verify that these priors are adequate to represent diffiuse or vague beliefs. In particular, the implied prior for the identified regression coefficient, $\widetilde{\beta}$, is symmetric and puts prior mass over a very large range of possible values. An approach which uses priors on the non-identified space to induce priors over the identified parameters is very useful in situations where a diffuse prior is desired. This approach could become cumbersome for situations in which very informative priors are desired. However, in situations in which informative priors are needed may require priors which are not conjugate and this renders the entire Gibbs sampler proposed unuseable. For

example, if you have strong prior information about the correlations between some, but not all, of the choice errors, this information will be impossible to encode using standard conjugate Inverted Wishart priors.

Given these conditionally conjugate priors, a standard Gibbs sampler can be constructed as in McCulloch and Rossi (1994).

$$\begin{aligned} w_i|\beta,\Sigma,y_i,X_i^d \quad & i=1,\ldots,n \\ \beta|\Sigma,w & \\ \Sigma|\beta,w & \end{aligned} \tag{3.5}$$

We note that $\{w_i\}, i=1,\ldots,n$, are independent given the data and other model parameters. The conditional posterior for w_i is a $p-1$ dimensional normal truncated to a cone. The insight of Geweke (1991) and McCulloch and Rossi (1994) is to create a Gibbs sampler that breaks the draw of w_i into a sequence of $p-1$ univariate truncated normal draws.

$$\begin{aligned} w_{ij}|w_{i,-j} \quad &\sim \quad N\left(\mu_{ij},\tau_{jj}^2\right) \\ &\times \quad \left[I(j=y_i)\,I(w_{ij}>max\left(w_{i,-j},0\right))+I(j\neq y_i)\,I(w_{ij}<max\left(w_{i,-j},0\right))\right] \end{aligned} \tag{3.6}$$

The moments of the truncated normal above are given in Rossi, Allenby, and McCulloch (2005) and the Gibbs sampler is implemented in the *bayesm* function, `rmnpGibbs`.

It is possible to put a prior directly on the identified parameters, see McCulloch, Polson, and Rossi (2000). However, the Gibbs sampler based on this prior is more highly autocorrelated. Imai and van Dyk (2005) propose a modified prior for the unidentified parameter space that has similar autocorrelation properties as the original McCulloch and Rossi (1994) paper but is less sensitive to extreme initial conditions.

3.2 Multivariate Probit Model

The binary probit model specifies a single binary outcome as a function of observed covariates. The logical extension of this model is the multivariate probit model which produces a p dimensional binary outcome. Classic examples of the multivariate probit model include

purchases in multiple categories as in Manchanda, Ansari, and Gupta (1999). In this application, households are observed to purchase in related categories of products, e.g. pasta and pasta sauce, and there is the possibility that there are correlations between the latent "attractiveness" or utility of related categories. Another common situation arises in survey market research where respondents can select as 0 to p items from a list of p items. Typically, these questions provide lists of products and the respondent is asked which products she purchases on a regular basis. Again the choices of items can be viewed as revealing an underlying pattern of similarity between products.

The multivariate probit is based on a latent p-variate normal regression model. The censoring mechanism is that we only observe the sign of the latent response vector.

$$\begin{aligned} w_i &= X_i\beta + \epsilon_i \quad \epsilon_i \sim N(0, \Sigma) \\ y_{ij} &= \begin{cases} 1 & \text{if } w_{ij} > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{3.7}$$

Edwards and Allenby (2003) view this latent representation as the problem of conducting a standard multivariate analysis of binary outcome data. Clustering in the "1" outcome in the observed dependent variable (conditional on covariates) reveals a correlation between items. The identification problem in the multivariate probit is that arbitrary scaling of the underlying latent variables is possible without changing the outcome variable. This means that in many applications only the correlation structure of the latent variables is identified. Chib and Greenberg (1998) present a MCMC algorithm for the identified space of parameters; however, this algorithm requires tuning and has not been shown to work in problems with a high-dimensional covariance matrix. The algorithm of Edwards and Allenby (2003) can work in 20+ dimensional problems with no tuning.

The identification problem in the multivariate probit depends on the structure of the X array. In the general case, X will include intercepts for each of the p choice alternatives and

covariates that are allowed to have different coefficients for each of the p outcomes.

$$X_i = \left(z_i' \otimes I_p\right)$$

z_i is a $d \times 1$ vector of observations on the covariates. Thus, X is a $p \times k$ matrix and $k = p \times d$. We interpret β as a stacked vector of coefficients representing the impact of each of the variables in z on the p latent outcomes.

$$\beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_d \end{bmatrix}$$

β_d represents the impact of z_{id} on the mean latent utilities for each of the p possible outcomes. If z is measured income, then there can be different effects of income on the probability of each of the p outcomes. We might expect that this would be true for a category with a wide range of product qualities. Higher income consumers might have a greater probability of purchasing the higher quality products. For this X structure, the identification problem is most severe. Each element of the latent w vector can be scaled by a different positive constant without altering the binary outcomes. Thus, the identified parameters for this model are the correlation matrix of the errors and appropriately scaled regression coefficients. That is, we can define a transformation from the unidentified to identified parameters by

$$\begin{aligned} \tilde{B} &= \Lambda B \\ R &= \Lambda \Sigma \Lambda \end{aligned} \tag{3.8}$$

where

$$
\begin{aligned}
B &= [\beta_1, \ldots, \beta_p] \\
\Lambda &= \begin{bmatrix} 1/\sqrt{\sigma_{11}} & & \\ & \ddots & \\ & & 1/\sqrt{\sigma_{pp}} \end{bmatrix}.
\end{aligned}
$$

If the coefficients on a given covariate are restricted to be equal across all p choices, then there are fewer unidentified parameters. We cannot scale each equation by a *different* positive constant. For example, if z were to contain an attribute of a set of product offerings, then we might consider restricting the impact of variation in this attribute to be the same across outcomes or impose the restriction, $\beta_{j1} = \ldots = \beta_{jp}$ for covariate j. In this case, we cannot scale each element of the latent vector by different scale factor and the identification problem is identical to the one in the Multinomial Probit Model.

Edwards and Allenby (2003) implement a Gibbs sampler for the multivariate Probit model using the non-identified approach in which a proper prior is imposed on the non-identified parameter space and only the posterior distribution of identified parameters (3.8) is reported. The Gibbs sampler is the same as in MNP model (3.5) except that the truncation points for the draw of the latent utilities (as in 3.6) are 0 as in the binary probit. This sampler is implemented in the ***bayesm*** function, `rmvpGibbs`. Some care should be exercised in the prior setting for the IW prior on Σ. Very diffuse priors on Σ are informative on the prior distribution of correlation coefficients. In particular, very diffuse priors imply a "U" shaped prior distribution for the correlation coefficients.

3.3 Count Regression Models

In some marketing applications, the outcome variable is best viewed as a count variable with potentially a large number of values. For example, if we observe physicians writing prescriptions for a specific drug in a given period of time, we might observe zeroes but also rather large numbers as well. Marketing actions such as visits by salespeople presumably increase

the expected number of prescriptions. Given the truncation at zero and the integer aspects of the data, standard regression models are not appropriate. The Poisson regression model is often used for count data but suffers from the restrictive assumption that the conditional mean and variance are required to be the same. In many applications, the data exhibit "over-dispersion," that is, the conditional variance exceeds the conditional mean. The negative binomial regression model provides a natural generalization of the Poisson model in the sense that it provides for any degree of over dispersion (see also Morrison and Schmittlein, 1988).

$$Pr\left(y_i = k|\lambda_i, \alpha\right) = \frac{\Gamma\left(\alpha + k\right)}{\Gamma\left(\alpha\right)\Gamma\left(k+1\right)}\left(\frac{\alpha}{\alpha+\lambda_i}\right)^{\alpha}\left(\frac{\lambda_i}{\alpha+\lambda_i}\right)^{k} \tag{3.9}$$

In this parameterization of the negative-binomial, λ is the mean and α is the over-dispersion parameter (as $\alpha \to \infty$, the negative binomial approaches the Poisson distribution). A standard log-link function provides a way of incorporating covariates.

$$ln\left(\lambda_i\right) = x_i'\beta \tag{3.10}$$

Priors for β and α can be conveniently chosen as

$$\begin{gathered} p\left(\beta, \alpha\right) = p\left(\beta\right)p\left(\alpha\right) \\ \beta \sim N\left(\underline{\beta}, \underline{A}^{-1}\right) \\ \alpha \sim G\left(\underline{a}, \underline{b}\right) \end{gathered} \tag{3.11}$$

An MCMC algorithm for the model in (3.9), (3.10), and (3.11) can be defined by a "Gibbs-style" combination of two Random Walk Metropolis methods to draw first from the posterior of $\beta|\alpha$ and then from the univariate posterior of $\alpha|\beta$. This sampler is implemented in the *bayesm* function, `rnegbinRw`. The random walk increment density can be chosen based on an estimated Hessian evaluated at the posterior mode or MLE. The *bayesm* function uses the Roberts and Rosenthal (2001) scaling suggestion.

4 Hierarchical Models

A fundamental premise of marketing is that customers differ both in preferences for product features as well as their sensitivities to marketing variables. Observable characteristics such as psycho-demographics can only be expected to explain a limited portion of the variation in tastes and responsiveness. Disaggregate data is required in order to measure customer heterogeneity. Typically, disaggregate data are obtained for a relatively large number of cross-sectional units but with a relatively short history of activity. In the consumer packaged goods industry, store level panel data are common, especially for retailers. There is also increased availability of customer level purchase data from specialized panels of consumers or from detailed purchase histories assembled from firm records. As the level of aggregation decreases, discrete features of sales data become magnified. The short time span of panel data coupled with the comparatively sparse information in discrete data means that we are unlikely to have a great deal of sample information about any one cross-sectional unit. If inference about unit-level parameters is important, then the prior will matter and that assessing informative priors will be important. Increasingly firms want to make decentralized marketing decisions that exploit more detailed disaggregate information. Examples include store or zone level pricing, targeted electronic couponing, and sales force activities in the pharmaceutical industry. This contrasts markedly with applications in micro-economics where the average response to a variable is often deemed more important. However, even the evaluation of policies which are uniform across some set of consumers will require information about the distribution of preferences in order to evaluate the effect on social welfare.

From a Bayesian perspective, modeling panel data is about the choice of a prior over a high dimensional parameter space. The hierarchical approach is one convenient way of specifying the joint prior over unit-level parameters. Clearly, this prior will be informative and must be in order to produce reasonable inferences. However, it is reasonable to ask for flexibility in the form of this prior distribution. In this section, we will introduce hierarchical models for general unit level models. Recognizing the need for flexibility in the prior, we will expand the set of priors to include mixtures of normal distributions.

4.1 A Generic Hierarchical Approach

Consider a cross-section of H units, each with a likelihood, $p\left(y_h|\theta_h\right), h = 1, \ldots, H$. θ_h is a $k \times 1$ vector. y_h generically represents the data on the hth unit and θ_h is a vector of unit-level parameters. While there is no restriction on the model for each unit, common examples include a multinomial logit or standard regression model at the unit level. The parameter space can be very large and consists of the collection of unit level parameters, $\{\theta_h, h = 1, \ldots, H\}$. Our goal will be to conduct a posterior analysis of these joint set of parameters. It is common to assume that units are independent conditional on θ_h. More generally, if the units are *exchangeable* (see Bernardo and Smith, 1994), then we require a prior distribution which is the same no matter what the ordering of the units are. In this case, we can write down the posterior for the panel data as

$$p\left(\theta_1, \ldots, \theta_H|y_1, \ldots, y_H\right) \propto \prod_{h=1}^{H} p\left(y_h|\theta_h\right) p(\theta_1, \ldots, \theta_H|\tau) \tag{4.1}$$

τ is a vector of prior parameters. The prior assessment problem posed by this model is daunting as it requires specifying a potentially very high dimensional joint distribution. One simplification would be to assume that the unit-level parameters are independent and identically distributed, *a priori*. In this case, the posterior factors and inference can be conducted independently for each of the H units.

$$p\left(\theta_1, \ldots, \theta_H|y_1, \ldots, y_H\right) \propto \prod_{h=1}^{H} p\left(y_h|\theta_h\right) p(\theta_h|\tau) \tag{4.2}$$

Given τ, the posterior in (4.2) is the Bayesian analogue of the classical *fixed effects* estimation approach. However, there are still advantages to the Bayesian approach in that an informative prior can be used. The informative prior will impart important shrinkage properties to Bayes estimators. In situations in which the unit-level likelihood may not be identified, a proper prior will regularize the problem and produce sensible inferences. The real problem is a practical one in that some guidance must be provided for assessing the prior parameters, τ.

The specification of the conditionally independent prior can be very important due to the scarcity of data for many of the cross-sectional units. Both the form of the prior and the values of hyperparameters are important and can have pronounced effects on the unit-level inferences. For example, consider a normal prior, $\theta_i \sim N\left(\bar{\theta}, V_\theta\right)$. Just the use of a normal prior distribution is highly informative regardless of the value of hyperparameters. The thin tails of the prior distribution will reduce the influence of the likelihood when the likelihood is centered far away from the prior. For this reason, the choice of the normal prior is far from innocuous. For many applications, the shrinkage of outliers is a desirable feature of the normal prior. The prior results in very stable estimates but at the same time this prior might mask or attenuate differences in consumers. It will, therefore, be important to consider more flexible priors.

If we accept the normal form of the prior as reasonable, a method for assessing the prior hyperparameters is required (Allenby and Rossi, 1999). It may be desirable to adapt the shrinkage induced by use of an informative prior to the characteristics of both the data for any particular cross-sectional unit as well as the differences between units. Both the location and spread of the prior should be influenced by both the data and our prior beliefs. For example, consider a cross-sectional unit with little information available. For this unit, the posterior should shrink toward some kind of "average" or representative unit. The amount of shrinkage should be influenced both by the amount of information available for this unit as well as the amount of variation across units. A hierarchical model achieves this result by putting a prior on the common parameter, τ. The hierarchical approach is a model specified by a sequence of conditional distributions, starting with the likelihood and proceeding to a two-stage prior.

$$\begin{aligned} &p\left(y_h|\theta_h\right) \\ &p\left(\theta_h|\tau\right) \\ &p\left(\tau|\underline{h}\right) \end{aligned} \tag{4.3}$$

The prior distribution on $\theta_h|\tau$ is sometimes called the first stage prior. In non-Bayesian applications, this is often called a random effect or random coefficient model and is regarded as

part of the likelihood. The prior on τ completes the specification of a joint prior distribution on all model parameters.

$$p\left(\theta_1, \ldots, \theta_H, \tau | \underline{h}\right) = p\left(\theta_1, \ldots, \theta_H | \tau\right) p\left(\tau | h\right) = \prod_{h=1}^{H} p\left(\theta_h | \tau\right) p\left(\tau | \underline{h}\right) \tag{4.4}$$

One way of regarding the hierarchical model is just as a device to induce a joint prior on the unit-level parameters, that is we can integrate out τ to inspect the implied prior.

$$p\left(\theta_1, \ldots, \theta_H | \underline{h}\right) = \int \prod_{h=1}^{H} p\left(\theta_h | \tau\right) p\left(\tau | \underline{h}\right) d\tau \tag{4.5}$$

It should be noted that, while $\{\theta_h\}$ are independent conditional on τ, the implied joint prior can be highly dependent, particularly if the prior on τ is diffuse (note: it is sufficient that the prior on τ should be proper in order for the hierarchical model to specify a valid joint distribution). To illustrate this, consider a linear model, $\theta_h = \tau + v_h$. τ acts as common variance component and the correlation between any two θs is

$$Corr\left(\theta_h, \theta_k\right) = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_v^2}$$

As the diffusion of the distribution of τ relative to v increases, this correlation tends toward one.

4.2 MCMC Schemes

Given the independence of the units conditional θ_h, all MCMC algorithms for a hierarchical models will contain two basic groups of conditional distributions.

$$\begin{gathered} p\left(\theta_h | y_h, \tau\right) \quad h = 1, \ldots, H \\ p\left(\tau | \left\{\theta_h\right\}, \underline{h}\right) \end{gathered} \tag{4.6}$$

As is well-known, the second part of this scheme exploits the conditional independence of y_h and τ. The first part of (4.6) is dependent on the form of the unit-level likelihood, while the

second part depends on the form of the first stage prior. Typically, the priors in the first and second stages are chosen to exploit some sort of conjugacy and the $\{\theta_h\}$ are treated as "data" with respect to the second stage.

4.3 Fixed Vs. Random Effects

In classical approaches, there is a distinction made between a "fixed effects" specification in which there are different parameters for every cross-sectional unit and random effects models in which the cross-sectional unit parameters are assumed to be draws from a super-population. Advocates of the fixed effects approach explain that the approach does not make any assumption regarding the form of the distribution or the independence of random effects from included covariates in the unit-level likelihood. The Bayesian analogue of the fixed effects classical model is an independence prior with no second-stage prior on the random effects parameters as in (4.2). The Bayesian hierarchical model is the Bayesian analogue of a random effects model. The hierarchical model assumes that each cross-sectional unit is exchangeable (possibly conditional on some observable variables). This means that a key distinction between models (Bayesian or classical) is what sort of predictions could be made for a new cross-sectional unit. In either the classical or Bayesian "fixed effects" approach, no predictions can be made about a new member of the cross-section as there is no model linking units. Under the random effects view, all units are exchangeable and the predictive distribution for the parameters of a new unit is given by

$$p\left(\theta_{h*}|y_1,\ldots,y_H\right) = \int p\left(\theta_{h*}|\tau\right) p\left(\tau|y_1,\ldots,y_H\right) d\tau \tag{4.7}$$

4.4 First Stage Priors

Normal Prior

A straightforward model to implement is a normal first stage prior with possible covariates.

$$\theta_h = \Delta' z_h + v_h, \quad v_h \sim N\left(0, V_\theta\right) \tag{4.8}$$

z_h is a $d \times 1$ vector of observable characteristics of the cross-sectional unit. Δ is a $d \times k$ matrix of coefficients. The specification in (4.8) allows the mean of each of the elements of θ_h to depend on the z vector. For ease of interpretation, we find it useful to subtract the mean and use an intercept.

$$z_h = (1, x_h - \bar{x})$$

In this formulation, the first row of Δ can be interpreted as the mean of θ_h.

(4.8) specifies a multivariate regression model and it is convenient, therefore, to use the conjugate prior for the multivariate regression model.

$$\begin{aligned} V_\theta &\sim IW\left(\underline{V}, \underline{\nu}\right) \\ \delta = vec\left(\Delta\right) | V_\theta &\sim N\left(\underline{\bar{\delta}}, V_\theta \otimes \underline{A}^{-1}\right) \end{aligned} \tag{4.9}$$

$\underline{A}$ is a $d \times d$ precision matrix. This prior specification allows for direct one-for-one draws of the common parameters, δ and V_θ. In `bayesm`, these draws can be achieved using the utility function, `rmultireg`.

Mixture of Normals Prior

While the normal distribution is flexible, there is no particular reason to assume a normal first-stage prior. For example, if the observed outcomes are choices among products, some of the coefficients might be brand specific intercepts. Heterogeneity in tastes for a product might be more likely to assume the form of clustering by brand. That is, we might find "clusters" of consumers who prefer specific brands over other brands. The distribution of tastes across consumers might then be multi-modal. We might want to shrink different groups of consumers in different ways or shrink to different group means. A multi-modal distribution will achieve this goal. For other coefficients such as a price sensitivity coefficient, we might expect a skewed distribution centered over negative values. Mixtures of multivariate normals are one way of achieving a great deal of flexibility (see, for example, Griffin, Quintana, and Steel (2010) and the references therein). Multi-modal, thick-tailed and skewed distributions are easily achieved from mixtures of a small number of normal components. For larger numbers

of components, virtually any joint continuous distribution can be approximated. The mixture of normals model for the first-stage prior is given by

$$\begin{aligned} \theta_h &= \Delta' z_h + v_h \\ v_h &\sim N\left(\mu_{ind}, \Sigma_{ind}\right) \\ ind &\sim MN\left(\pi\right) \end{aligned} \tag{4.10}$$

π is a $K \times 1$ vector of multinomial probabilities. This is a latent version of a mixture of K normals model in which a multinomial mixture variable, denoted here by ind, is used. In the mixture of normal specification, we remove the intercept term from z_h and allow v_h to have a non-zero mean. This allows the normal mixture components to mix on the means as well as on scale, introducing more flexibility. As before, it is convenient to demean the variables in z. A standard set of conjugate priors can be used for the mixture probabilities and component parameters, coupled with a standard conjugate prior on the Δ matrix.

$$\begin{aligned} \delta = vec\left(\Delta\right) &\sim N\left(\underline{\bar{\delta}}, \underline{A}_{\delta}^{-1}\right) \\ \pi &\sim D\left(\underline{\alpha}\right) \\ \mu_k &\sim N\left(\underline{\bar{\mu}}, \Sigma_k \otimes \underline{a}_{\mu}^{-1}\right) \\ \Sigma_k &\sim IW\left(\underline{V}, \underline{\nu}\right) \end{aligned} \tag{4.11}$$

Assessment of these conjugate priors is relatively straight-forward for diffuse settings. Given that the θ vector can be of moderately large dimension (>5) and the θ_h parameters are not directly observed, some care must be exercised in the assessment of prior parameters. In particular, it is customary to assess the Dirichlet portion of the prior by using the interpretation that the $K \times 1$ hyperparameter vector, $\underline{\alpha}$, is an observed classification of a sample of size, $\sum \underline{\alpha}_k$, into the K components. Typically, all components in $\underline{\alpha}$ are assessed equal. When a large number of components are used, the elements of α should be scaled down in order to avoid inadvertently specifying an informative prior with equal prior probabilities on a large number of components. We suggest a setting of $\underline{\alpha}_k = .5/K$.

As in the single component normal model, we can exploit the fact that, given the $H \times k$ ma-

trix, Θ , whose columns consist of each θ_h values and standard conditionally conjugate priors in (4.11), the mixture of normals model in (4.10) is easily handled by a standard unconstrained Gibbs sampler which includes augmentation to include the latent vector of component indicators (see Rossi, Allenby, and McCulloch (2005), section 5.5.1). The latent draws can be used for clustering as discussed below. We should note that any label-invariant quantity such as a density estimate or clustering is not affected by the "label-switching" identification problem (see Fruhwirth-Schnatter (2006) for a discussion). In fact, the unconstrained Gibbs sampler is superior to various constrained approaches in terms of mixing.

A tremendous advantage of Bayesian methods when applied to mixtures of normals is that, with proper priors, Bayesian procedures do not overfit the data and provide reasonable and smooth density estimates. In order for a component to obtain appreciable posterior mass, there must be enough structure in the "data" to favor the component in terms of a Bayes factor. As is standard in Bayesian procedures, the existence of a prior puts an implicit penalty on models with a larger number of components. It should also be noted that the prior for the mixture of normals puts positive probability on models with less than K components. In other words, this is really a prior on models of different dimensions. In practice, it is common for the posterior mass to be concentrated on a set of components of much smaller size than K.

The posterior distribution of any ordinate of the joint (or marginal densities) of the mixture of normals can be constructed from the posterior draws of component parameters and mixing probabilities. In particular, a Bayes estimate of a density ordinate can be constructed.

$$\hat{d}(\theta) = \frac{1}{R}\sum_{r=1}^{R}\sum_{k=1}^{K}\pi_k^r \phi(\theta|\mu_k^r, \Sigma_k^r) \tag{4.12}$$

Here the superscript r refers to an MCMC posterior draw and $\phi(\cdot)$ the k-variate multivariate normal density. If marginals of sub-vectors of θ are required, then we simply compute the required parameters from the draws of the joint parameters. The *bayesm* plot method for normal mixtures, `plot.bayesm.nmix`, will automatically compute and plot univariate and

bivariate marginals for general normal mixtures.

Dirichlet Process Priors

While it can be argued that a finite mixture of normals is a very flexible prior, it is true that the number of components must be pre-specified by the investigator. Given that Bayes methods are being used, a practical approach would be to assume a very large number of components and allow the proper priors and natural parsimony of Bayes inference to produce reasonable density estimates. For large samples, it might be reasonable to increase the number of components in order accommodate greater flexibility. The Dirichlet Process (DP) approach can, in principle, allow the number of mixture components to be as large as the sample size and potentially increase with the sample size. This allows for a claim that a DP prior can facilitate general non-parametric density estimation. Griffin, Quintana, and Steel (2010) provides a discussion of the DP process approach to density estimation. We review only that portion of this method necessary to fix notation for use within a hierarchical setting.

Consider a general setting in which each θ_h is drawn from a possibly different multivariate normal distribution.

$$\theta_h \sim N\left(\mu_h, \Sigma_h\right)$$

The DP process prior is a hierarchical prior on the joint distribution of $\{(\mu_1, \Sigma_1), \ldots, (\mu_H, \Sigma_H)\}$. The DP prior has the effect of grouping together cross-section units with the same value of (μ, Σ) and specifying a prior distribution for these possible "atoms."

The DP process prior is denoted $G\left(\alpha, G_0\left(\lambda\right)\right)$. $G\left(\cdot\right)$ specifies a distribution over distributions that is centered on the base distribution, G_0, with tightness parameter, α. Under the DP prior, G_0 is the marginal prior distribution for the parameters for any one cross-sectional unit. α specifies the prior distribution on the clustering of units to a smaller number of unique (μ, Σ) values. Given the normal base distribution for the cross-sectional parameters, it is convenient to use a natural conjugate base prior.

$$G_0\left(\lambda\right): \quad \mu_h|\Sigma_h \sim N\left(\underline{\bar{\mu}}, \frac{1}{\underline{a}} \times \Sigma_h\right), \quad \Sigma_h \sim IW\left(\underline{V}, \underline{\nu}\right) \tag{4.13}$$

λ is the set of prior parameters in (4.13), $\underline{\overline{\mu}}, \underline{a}, \underline{\nu}, \underline{V}$.

In our approach to a DP model, we also put priors on the DP process parameters, α and λ. The Polya Urn representation of the DP model can be used to motivate the choice of prior distributions on these process parameters. α influences the number of unique values of (μ, Σ) or the probability that a new set of parameter values will be "proposed" from the base distribution, G_0. λ governs the distribution of proposed values. For example, if we set λ to put high prior probability on small values of Σ, then the DP prior will attempt to approximate the density of parameters with normal components with small variance. It is also important that the prior on μ put support on a wide enough range of values to locate normal components at wide enough spacing to capture the structure of the distribution of parameters. On the other hand, if we set very diffuse values of λ then this will reduce the probability of the "birth" of a new component via the usual Bayes Factor argument.

α induces a distribution on the number of distinct values of (μ, Σ) as shown in Antoniak (1974).

$$Pr\left(I^* = k\right) = \left\| S_n^{(k)} \right\| \alpha^k \frac{\Gamma\left(\alpha\right)}{\Gamma\left(n+\alpha\right)} \tag{4.14}$$

$S_n^{(k)}$are Sterling numbers of the first kind. I^* is the number of clusters or unique values of the parameters in the joint distribution of $(\mu_h, \Sigma_h, h = 1, \ldots, H)$. It is common in the literature to set a Gamma prior on α. Our approach is to propose a simple and interpretable distribution for α.

$$p\left(\alpha\right) \propto \left(1 - \frac{\alpha - \underline{\alpha}_l}{\underline{\alpha}_u - \underline{\alpha}_l}\right)^{\underline{\phi}} \tag{4.15}$$

$\alpha \in (\underline{\alpha}_l, \underline{\alpha}_u)$. We assess the support of α by setting the expected minimum and maximum number of components, I^*_{min} and I^*_{max}. We then invert to obtain the bounds of support for α. It should be noted that this device does not restrict the support of the number of components but merely assesses an informative prior that puts most of the mass of the distribution of α on values which are consistent with the specified range in the number of unique components. A draw from the posterior distribution of α can easily be accomplished as I^* is sufficient and we can use a griddy Gibbs sampler as this is simply a univariate draw.

Priors on λ (4.13) can be also be implemented by setting $\underline{\bar{\mu}} = 0$ and letting $V = \nu v I_k$. If $\Sigma \sim IW(\nu v I_k, \nu)$, then $mode(\Sigma) = \frac{\nu}{\nu+2} v I_k$. This parameterization helps separate the choice of a location for the Σ matrix (governed by v) from the choice of the tightness on the prior for Σ (ν). In this parameterization, there are three scalar parameters that govern the base distribution, (a, v, ν). We take them to be a priori independent with the following distributions.

$$\begin{aligned} p(a, v, \nu) &= p(a)\, p(v)\, p(\nu) \\ a &\sim U(\underline{a}_l, \underline{a}_u) \\ v &\sim U(\underline{v}_l, \underline{v}_u) \\ \nu &= dim(\theta_h) - 1 + \exp(z),\ z \sim U(\underline{z}_l, \underline{z}_u),\ z_l > 0 \end{aligned} \tag{4.16}$$

It is a simple matter to write down the conditional posterior given that the unique set of (μ, Σ) are sufficient. The set of I^* unique parameter values is denoted $\Delta^* \left\{ \left(\mu_i^*, \Sigma_j^* \right), j = 1, \ldots, I^* \right\}$. The conditional posterior is given by

$$\begin{aligned} p(a, v, \nu | \Delta^*) \propto & \prod_{i=1}^{I^*} \left| a^{-1} \Sigma_i^* \right|^{-1/2} \exp\left(-\frac{a}{2} (\mu_i^*)' (\Sigma_i^*)^{-1} \mu_i^* \right) \\ & |\nu v I_k|^{\nu/2} \left| \Sigma_i^* \right|^{-(\nu+k+1)/2} \operatorname{etr}\left(-\frac{1}{2} \nu v \Sigma_i^* \right) p(a, v, \nu) \end{aligned} \tag{4.17}$$

We note that the conditional posterior factors and that, conditional on Δ^*, a and (ν, v) are independent.

4.5 Examples

Linear Hierarchical Model

The linear hierarchical model with a normal first stage prior is covered in Koop (2003) and Geweke (2005). An implementation is available in the *bayesm* routine, `rhierLinearModel`. Given the modularity of the MCMC approach, it is a simple matter to extend the model to include finite mixtures of normals or DP process priors. An implementation for finite normal

mixtures is available in the *bayesm* routine, `rhierLinearMixture`.

Multinomial Logit Models

If the cross-sectional model is a multinomial logit (2.13), then finite mixtures of normals or DP process priors can be used. An implementation for finite normal mixtures is available in the *bayesm* routine, `rhierMnlMixture` and for DP process priors in `rhierMnlDP`.

Multivariate Ordinal Probit with Scale Usage Heterogeneity

Many surveys ask a battery of questions which are on a K point ratings scale. This data is ordinal and there is some question as to whether there is also interval information in the responses. Typically, a battery of M questions are administered to each of N respondents. The data for the survey can the represented as a $N \times M$ array, $X = [x_{ij}]$. An example of this sort of survey might be a customer satisfaction survey with M questions about service delivery or product quality. One useful approach to this problem is to assume that the responses are from an underlying M dimensional latent system, a natural generalization of the univariate ordinal probit.

Assume there are $K+1$ common and ordered cut-off points $\{c_k : c_{k-1} \leq c_k, k = 1, \ldots K\}$ with $c_0 = -\infty$, $c_K = \infty$, such that for all i, j, and k,

$$\begin{aligned} x_{ij} = k \quad &\text{if} \quad c_{k-1} \leq y_{ij} \leq c_k \\ y_i \quad &\sim \quad N\left(\mu_i^*, \Sigma_i^*\right) \end{aligned} \tag{4.18}$$

y_i is a $M \times 1$ vector. This can be interpreted as implying that each respondent has their own underlying covariate structure for the M survey questions. As a special case, all respondents have the same latent variable distribution and we would obtain a multivariate ordinal probit. We introduce respondent-specific parameters to accommodate what we have termed scale usage heterogeneity. That is, we see that some respondents have a tendency to use only a portion of the ratings scale. For example, some respondents tend to use either the middle, low end or high end of the scale. Experts in survey research have noted this for years and have

noted systematic cultural differences in scale usage as well. Scale usage heterogeneity obscures the relationships in the data and must be adjusted for. Inferences about scale usage are only possible if multiple questions are administered to the same respondent. Rossi, Gilula, and Allenby (2001) consider a hierarchical approach to incorporating scale usage heterogeneity.

Scale usage patterns can be created by a respondent-specific location and scale shift in the latent continuous variables. For example, a respondent who uses the top end of the scale can be modeled as someone who has a positive shift in the mean of the latent variables (across all questions) and has a lower variance. The location scale model is a parsimonious way of creating heterogeneity between respondents. The model does assume, however, that there is a meaningful common latent scale between all respondents.

$$y_i = \mu + \tau_i \iota + \sigma_i z_i, \quad z_i \sim N(0, \Sigma) \tag{4.19}$$

We employ a hierarchical model for the distribution of the location and scale parameters over respondents.

$$\begin{bmatrix} \tau_i \\ ln\sigma_i \end{bmatrix} \sim N(\phi, \lambda) \tag{4.20}$$

The scale usage translation must be restricted so that the model is identifed. Conditional on the cut-offs, we cannot allow for arbitrary translation and scaling of the latents. To avoid this identification problem, we set $E[\tau_i] = 0$ and $\text{Mode}(\sigma_i) = 1$. This can be achieved by specifying that

$$\begin{aligned} \phi_1 &= 0 \\ \phi_2 &= \lambda_{22} \end{aligned} \tag{4.21}$$

Even with these restrictions to achieve identification of the (τ_i, σ_i) parameters, there are still identification restrictions that must be imposed on the cut-off parameters. The cut-off parameters exhibit a location and scale invariance problem. We fix the sum of the cut-offs

and the sum of the squared cut-offs to remove the location/scale invariance problem.

$$\begin{aligned} \sum_k c_k &= m_1 \\ \sum_k c_k^2 &= m_2 \end{aligned} \tag{4.22}$$

In order to reduce the number of cut-off parameters, particularly for 7 or larger point scales, we introduce a quadratic parameterization of the cut-off parameters.

$$c_k = a + bk + ek^2 \tag{4.23}$$

Given the parameterization in (4.23) and the identification restrictions (4.22), there is only one free parameter, e.

The priors for this model are given by

$$\begin{gathered} \pi(\mu, \Sigma, \phi, \Lambda, e) = \pi(\mu)\,\pi(\Sigma)\,\pi(\phi)\,\pi(\Lambda)\,\pi(e) \\ \pi(\mu) \propto \text{constant} \\ e \sim U(-.2, .2) \\ \Sigma \sim IW(\underline{V}_\Sigma, \underline{\nu}_\Sigma) \\ \Lambda \sim IW(\underline{V}_\Lambda, \underline{\nu}_\Lambda) \end{gathered} \tag{4.24}$$

The range of the prior on e is chosen to provide an adequate range of possible patterns of cut-offs. We note that given the identification restrictions (4.21), the prior on Λ induces a prior on ϕ. Rossi, Gilula, and Allenby (2001) define a Gibbs sampler for this model that uses a method of collapsing for acceleration by integrating out the latents, $\{y_i\}$, from one of the draws (see the appendix for details). An implementation is available in the ***bayesm*** routine, `rscaleUsage`.

Computational Notes

Customized Metropolis-Hastings for the General Hierarchical Problem In the generic hierarchical model (4.3), a likelihood is postulated for each cross-sectional "unit" and this is coupled with a two-stage hierarchical prior. If a Metropolis-Hastings random walk step is used for the draw of $\theta_h|y_h, \tau, h$, then, in principle, any model can be used for the unit likelihood. Given the flexibility of mixture of normals priors and Dirichlet Process priors, a Gibbs sampler can usually be constructed for the draws of $\tau|\{\theta_h\}$.

The only limitation is that the M-H random walk methods only work well if the random walk increments can be tuned to conform as closely as possible to the curvature in the conditional posterior

$$p(\theta_h|y_h, \tau) \propto p(y_h|\theta_h)\, p(\theta_h|\tau) \tag{4.25}$$

Therefore, for all except the most regular models, it will be necessary to customize the Metropolis chains for each cross-sectional unit. Without prior information on highly probable values of the first stage prior parameters, τ, it will be difficult to use the strategy of trial runs to tune the Metropolis chains given that a large fraction of cross-sectionals have limited information about the model parameters. One other possibility that is often employed is to use the pooled likelihood for all units and scale the Hessian from this pooled likelihood for the number of observations in any one unit (see Allenby and Rossi, 1993). Define $\bar{\ell}(\theta) = \prod_{h=1}^{H} \ell(\theta_h|y_h)$ as the pooled likelihood. The scaled Hessian is given by

$$\bar{H}_h = \frac{n_h}{N} \frac{\partial^2 log\bar{\ell}}{\partial\theta_h \partial\theta_h'}|_{\theta=\hat{\theta}_{MLE}} \tag{4.26}$$

$N = \sum_{h=1}^{H} n_h$. n_h is the number of observations for cross-sectional unit h. The scaled Hessian is a curvature estimate for each cross-sectional unit but it is based on a mixture across units. While this will get the scaling or units approximately correct for each element of θ, there is no guarantee that this curvature estimate will approximate the correlation structure in each individual unit. The virtue of the use of a Hessian based on the pooled sample is that the pooled MLE is often easy to find and has a non-singular Hessian.

At the opposite extreme from the use of the pooled MLE would be to use Hessian estimates constructed from each unit likelihood. This would require that the MLE exist and that the Hessian is non-singular for each cross-sectional unit likelihood. For choice model applications, this would require, at a minimum, that each cross-sectional unit be observed to choose at least once from all choice alternatives (sometimes termed a "complete" purchase history). If a unit does not chose a particular alternative and if an alternative-specific intercept is included in the model, then the MLE will not be defined for this unit. There would exist a direction of recession in which an intercept will drift off to $-\infty$ with an increasing likelihood. What is required is a regularization of the unit-level likelihood for that sub-sample of units with singular Hessians or non-existent MLEs. Our proposal is to borrow from the "fractional" likelihood literature for the purpose of computing an estimate of the unit level Hessian. This is only used for the Random Walk Metropolis increment covariance matrix and is *not* used to replace the unit level likelihood in posterior computations.

To compute the Hessian, we form a fractional combination of the unit-level likelihood and the pooled likelhood.

$$\ell_h^* (\theta) = \ell_h (\theta)^{(1-w)} \bar{\ell} (\theta)^{w\beta} \tag{4.27}$$

The fraction, w, should be chosen to a rather small number so that only a "fraction" of the pooled likelihood, $\bar{\ell}$, is combined with the unit likelihood, ℓ_h, to form the regularized likelihood. β is chosen to properly scale the pooled likelihood to the same order as the unit likelihood. $\beta = \frac{n_h}{N}$. (4.27) is maximized to estimate the Hessian at the "modified" MLE. This Hessian can be combined with the normal covariance matrix from the unit-level conditionally normal prior (note: if the prior is of the mixture of normal form, we are conditioning on the indicator for this unit). If the RW Metropolis increments are $N\left(0, s^2\Omega\right)$, then

$$\Omega = \left(H_h + V_\theta^{-1}\right)^{-1} \tag{4.28}$$

$$H_h = -\frac{\partial^2 log \ell_h^*}{\partial\theta\partial\theta'}|_{\theta=\hat{\theta}_h} \tag{4.29}$$

$\hat{\theta}_h$ is the maximum of the modified likelihood in (4.27). This customized MH method is

illustrated in the *bayesm* routine, `rhierMnlRwMixture`.

Clustering Using a Mixture of Normals In many marketing applications, some sort of clustering method is desired to group "observations" or units that exhibit some sort of similarity. Most clustering methods are based on distance metrics that are related to the normal distribution. If a mixture of normals is used, a very general clustering method can be developed. "Observations" are grouped into various normal sub-populations. The only caveat is that there is no restriction that variance of each normal mixture component be "smaller" than the variation across components. For example, a thick-tailed distribution across units can be approximated by the mixture of a small variance normal with a very large variance normal. Observations that get clustered into the large variance component should be interpreted as "similar" only to the extent that they are outliers.

In addition to the metric by which similarity is gauged, there is also a question as to what variables should be used as the basis of clustering. Traditional methods cluster units on the basis of observables such as psycho-graphics. However, if unit level behavioral data is available, it is now possible to cluster units on the basis of unit-level parameters. In some applications, these parameters can be interpreted as representing unit level tastes. This form of clustering on the basis of inference about the unit-level parameters, θ_h, can be termed "behavioral" clustering. Given that psycho-graphics are not very predictive of brand preferences or sensitivity to marketing variables, it is likely that behavioral clustering will be very useful in marketing applications.

To cluster based on a mixture of normals model (4.10), we use the latent indicators of component "assignment". We note that this can apply either when a mixture of normals approach is applied directly to data (density estimation) or as the random coefficient distribution. There can be a fixed number of normal components or a random number of components as in the DP models. All that is required is that we have draws from the posterior distribution of the indicator variables. These draws of the indicator variables, *ind*, can be used to form a

similarity matrix,

$$\begin{aligned} S &= [s_{i,j}] \\ s_{i,j} &= \begin{cases} 0 & if\ ind_i \neq ind_j \\ 1 & if\ ind_i = ind_j \end{cases} \end{aligned} \tag{4.30}$$

We note that the similarity matrix is invariant to label-switching. (4.30) defines a function from a given partition or classification of the observations to the similarity matrix. To emphasize this dependence, we will denote this function as, $S(ind)$. That is, for any clustering of the observations defined in an indicator vector, we can compute the associated similarity matrix. We can also find, for any similarity matrix, a classification or indicator vector consistent with the given similarity matrix. This function we denote by $ind = g(S)$.

By simply averaging over the draws from the marginal posterior of the indicator variables, we can estimate the posterior expectation of the similarity matrix.

$$\begin{aligned} S^* &= E_{ind|data}[S(ind)] \\ \hat{S^*} &= \frac{1}{R}\sum_{r=1}^{R} S(ind^r) \end{aligned} \tag{4.31}$$

Given the expected similarity matrix, the clustering problem involves the assignment or partition of the units so as to minimize some sort of loss function. Let ind be an assignment of units to groups and $L(S^*, S(ind))$ be a loss function, then we can define the clustering algorithm as the solution to the following problem:

$$min_{ind} L(S^*, S(ind)) \tag{4.32}$$

In general, this problem is a difficult optimization problem involving non-continuous choice variables. One could adopt two heuristics for the solution of the problem: 1. simply "classify" two observations as in the same group if the posterior expectation of similarity is greater than a cut-off value. 2. find the posterior draw which minimizes loss. A simple loss function would

be the sum of the absolute values of the differences between estimated posterior similarity and the implied similarity for a given value of the indicator or classification variable.

$$ind_{opt} = argmin_{\{ind^r\}} \left[\sum_i \sum_j \left| \hat{S}_{ij} - S\left(ind^r\right)[i,j] \right| \right] \tag{4.33}$$

The second heuristic uses the MCMC chain as a stochastic search process. The *bayesm* routine, `clusterMix`, implements both heuristics.

Data Structures for Panel Data and Normal Mixture Draws Both generic panel data and normal mixture MCMC output require data structures somewhat different from standard arrays. A generalized vector or list is an appropriate structure. That is, we require that the panel data be indexed by panelist or cross-sectional unit, but the data for each unit might consist of groups of objects of different types such as vectors and arrays. One might want to append information for customized MCMC draws to the panel data itself. This would facilitate retrieval for an MCMC method based on a hierarchical structure. Similarly, a set of draws from a normal mixture consists of a set of R objects each one of which is a set of objects of different type and dimension. In R, the list data structure is ideally suited for this purpose. A list is a generalized vector, each element of which can be any valid type of R object including a vector, array, function, or list itself. Lists can be nested, therefore, to any desired level. For example, it is possible to have a list of lists.

Panel data can be stored as a list of lists. In R notation, we could store a regression-style panel data set as the object `panel_data` which is a list of H panelists. The hth panelist would be indexed by `panel_data[[h]]` which is a list of two elements y, and X (the R syntax for indexing a list is the double square bracket). That is, `panel_data[[h]][[1]]`$=y_h$ and `panel_data[[h]][[2]]`$=X_h$. This approach avoids storage of indices into arrays or the use of ragged arrays. Given the pairwise list structure in R, it is a simple matter to add in new elements. For example, suppose you wish to add a customized Hessian (see (4.29)) to the panel data, is simple matter to define

`panel_data[[h]]=c(panel_data[[h]]],hess)`. "`c()`" is the R function for concatena-

tion.

The list structure is even more useful for storage of MCMC draws for normal components. A mixture of K multivariate normals can be represented as a list of K lists, `mix_norm[[k]][[1]]`=μ, `mix_norm[[k]][[2]]`=Σ (in practice, of course, we might store the inverse of the Cholesky root of Σ for ease of use in density evaluation). The MCMC draws would then be stored as a list of length R (R draws) of lists of K lists. This is the structure used in both DP and Finite Mixture functions in *bayesm*.

5 Non-random Marketing Mix Variables.

All of the models considered so far are motivated by a regression-style or conditional response model. That is, we model the distribution of some response variable (such as choice or sales) conditional on a set of covariates which include marketing mix variables.

$$\begin{array}{c} p\left(y|x,\theta\right) \\ p\left(x|\tau\right) \end{array} \tag{5.1}$$

Typically, we assume that the marginal distribution of the x variables is not related to the conditional distribution of $y|x$. However, we must recognize that the market mix variables in x are not set at random or independently of the sales response equation. Classic examples might include price setting where managers set prices on the basis of predictions of y. For example, suppose a retailer is aware that a manufacturer is going to issue a rebate or coupon in the next period and this will affect sales of an item. It is entirely possible that the retailer will set price taking this "demand shock" into account. However, the statistician observing price and sales data might not observe the coupon or rebate events and fail to take account of the strategic price-setting behavior. Suppose that the coupon drop made demand more in-elastic and the retailer raised prices in response. This could have the effect of making prices appear to have less effect on demand.

In the hierarchical setting, Manchanda, Rossi, and Chintagunta (2004) consider the situation in which the x variable is set with partial knowledge of the response parameter, θ, at the

unit level. The application considered is the allocation of sales force to various accounts. If the cost of sales force visits is roughly equal across accounts, then we would assume that the sales manager would allocate a budget-constrained sales force to the most responsive accounts. This means that the level of x is related to the cross-sectional unit value, θ_h.

One approach to the problem of strategically determined x values or non-random x is to model the choice of x as a function of the sales response equation as well as cost considerations of the firm. This would provide a model for the joint distribution of both x and y. The problem with this approach is that it is usually based on the assumption that the firm behaves optimally in the determination of the x values. This gives rise to an optimality conundrum in that if we assume optimality we can only estimate demand and firm parameters and have little to say to firms as to how to improve profitability. One way out of the optimality conundrum is to assume that firms set prices optimally with respect to some information set that does not have complete information about the model parameters. In this manner, improved profitability can be obtained from a richer information set (see, for example, Rossi, McCulloch, and Allenby (1996)). We will not take this approach here but instead consider the possibility that there exists what econometricians call an instrumental variable.

5.1 Bayesian Instrumental Variables

If we want to avoid making assumptions about precisely how the x variables are set, one approach is to assume that there is some portion of the variation in x that is *exogeneous* or determined by factors independent of y (analogous to true experimental variation). It will be useful to introduce the notion of an error term or driving variable to the determination of y. The sales response model now can be written:

$$y = f_y(x, \epsilon_y|\theta) \tag{5.2}$$

This equation is sometimes termed the "structural equation" but it does not represent the conditional distribution of $y|x$ as x is not independent of ϵ_y. We postulate the existence of an "instrumental variable" that is independent of ϵ_y. That is, we assume that x is driven in

part by the instrument and another error term that is correlated or dependent on ϵ_y.

$$x = f_x\left(z, \epsilon_x | \omega\right) \tag{5.3}$$

Classical instrumental variable methods merely exploit the fact that ϵ_y and z are assumed to be uncorrelated. In a Bayesian approach, a full-specified likelihood function must be used. Given a joint distribution on (ϵ_y, ϵ_x), we can derive the joint distribution of (y, x) which is the "reduced form."

$$p\left(x, y | z, \theta, \omega\right) \tag{5.4}$$

An important special case of (5.4) is the case of linear equations.

$$\begin{aligned} x &= \delta' z + \epsilon_x \\ y &= \beta x + \epsilon_y \end{aligned} \tag{5.5}$$

Note that, for ease of exposition, we have not included intercepts in the equation or other "exogenous" variables in the y equation (see Rossi, Allenby, and McCulloch (2005) for the general case). Our discussion includes the most common case in which there is only one potentially "endogenous" variable and an arbitrary number of instruments. We should note that while the first equation in (5.5) is a regression equation, the second equation is not if the error terms are dependent. That is, $p\left(\epsilon_y | x\right) \neq p\left(\epsilon_y\right)$ (see Lancaster, 2004).

Normal Errors

We will start our discussion of the linear instrumental variables model using bivariate normal errors. This will provide the basic intuition for Bayesian inference and can easily extended if the errors have a mixture of normals or DP process errors.

$$\begin{pmatrix} \epsilon_x \\ \epsilon_y \end{pmatrix} \sim N\left(0, \Sigma\right) \tag{5.6}$$

With normal errors, it is a simple matter to derive the likelihood function for the linear model. To derive the joint distribution of $y, x|z$, we simply substitute into the structural equation from the instrumental variables equation.

$$\begin{aligned} x &= \pi_x' z + v_x \\ y &= \pi_y' z + v_y \end{aligned} \tag{5.7}$$

with

$$\begin{pmatrix} v_x \\ v_y \end{pmatrix} \sim N(0, \Omega), \ \Omega = A\Sigma A', \ A = \begin{bmatrix} 1 & 0 \\ \beta & 1 \end{bmatrix} \tag{5.8}$$

and

$$\pi_x = \delta, \ \pi_y = \beta\delta.$$

Our view is that it is best to put priors directly on (β, δ, Σ) rather than on the reduced form coefficients and reduced form error covariance matrix. In particular, it would be inappropriate to assume that Ω and π_y are *a priori* independent since both sets of parameters depend on β. A useful starting point is to use conditionally conjugate and independent priors for the linear system.

$$\delta \sim N\left(\underline{\bar{\delta}}, \underline{A}_{\delta}^{-1}\right), \ \beta \sim N\left(\underline{\bar{\beta}}, \underline{a}_{\beta}^{-1}\right), \ \Sigma \sim IW\left(\underline{V}, \underline{\nu}\right) \tag{5.9}$$

It is easy to define a Gibbs sampler for the system (5.5) and (5.9). The Gibbs sampler contains three sets of conditional posterior distributions.

$$\begin{aligned} &\beta|\delta, \Sigma, y, x, Z \\ &\delta|\beta, \Sigma, y, x, Z \\ &\Sigma|\beta, \delta, y, x, Z \end{aligned} \tag{5.10}$$

The intuition for the Gibbs sampler is the same intuition that motivates the "endogeneity" problem in the first place. We know that the linear structural equation for y is not a valid regression equation because the error term as a non-zero mean which depends on x. However, the first distribution in the Gibbs sampler conditions on δ which means that we can "observe"

ϵ_x and the conditional distribution of $\epsilon_y|\epsilon_x$ can be derived. This distribution can be used to convert the structural equation into an equation with a $N(0,1)$ error term.

$$\left(y - \frac{\sigma_{xy}}{\sigma_x^2}\epsilon_x\right) = \beta x + u,\ u \sim N\left(0, \sigma_y^2 - \frac{\sigma_{xy}}{\sigma_x^2}\right) \tag{5.11}$$

Dividing through by σ_u converts the first Gibbs sampler draw into Bayes regression with a unit variance error term. The second conditional in the Gibbs sampler is simply a restricted two-variate regression which can be achieved easily by "doubling" the observations with rows of z_i and βz_i. The final Gibbs sampler conditional is a standard IW draw. This Gibbs sampler is implemented in the *bayesm* routine, `rivGibbs`.

Mixture of Normals and DP Errors

A legitimate concern about the Bayesian "IV" procedure is the additional specification of normal error terms in both the structural and "first-stage" instrument equation. This is not required for classical IV estimators. Usually, this is a concern about the possible inconsistency of an estimator based on a mis-specified likelihood. Equally important, but rarely appreciated, is the possibility for improved inference if it is possible to detect and model the non-normal structure in the error term. For example, suppose that the errors terms are a mixture of a normal error with small variance and another normal with a very large variance. The single component normal error model may be sensitive to the outliers and will treat the outlying error terms inefficiently. In principle, it should be possible to detect and down-weight observations that appear to have errors drawn from the outlying component. The classical IV approach sacrifices efficiency for the sake of consistency. In a semi-parametric Bayesian approach, it is theoretically possible to be robust to mis-specification while not reducing efficiency. That is, in the normal case we might not lose much or any efficiency but in the non-normal case we exploit the structure and construct an efficient procedure.

A logical place to start a departure from normality is the mixture of normals model. Intuitively, if we condition on the latent indicator variable for the membership in normal components, we should be able to reuse the Gibbs sampler for the normal model as we

can adjust and properly normalize for different variances. If we use mixture of normals to approximate any unknown bi-variate density of the error terms in the linear system (5.5), we need to be careful to allow the means of each component to be non-zero. A mixture of normals with zero or equal means is simply a scale mixture and cannot approximate skewed or multi-modal distributions. For this reason, we will allow non-zero means for the error terms and remove the intercepts from the model.

$$\begin{aligned} x &= Z\delta + \epsilon_x^* \\ y &= \beta x + \epsilon_y^* \\ \begin{pmatrix} \epsilon_x^* \\ \epsilon_y^* \end{pmatrix} &\sim N\left(\mu_{ind}, \Sigma_{ind}\right) \end{aligned} \tag{5.12}$$

To complete this model, we need to put a prior over the number of values normal components. In the standard finite mixture of normals, we assume that there up to K possible unique values. As discussed in Griffin, Quintana, and Steel (2010), a DP prior can be used which puts positive prior probability on up to N unique components. The DP process, $G\left(\alpha, G_0\right)$, defines this prior. We assess a prior on α (4.15) and directly set the hyper parameters for the base prior distribution, G_0 (see 4.13). We use the following parameterization of G_0.

$$G_0: \ \mu|\Sigma \sim N\left(0, \underline{a}^{-1}\Sigma\right), \ \Sigma \sim IW\left(\underline{c}I_2, \underline{\nu}\right) \tag{5.13}$$

To assess G_0, we center and scale both y and x. For centered and scaled dependent variables, we would expect that the bivariate distribution of the errors terms is concentrated on the region,$[-2,2] \times [-2,2]$. In order to achieve full flexibility, we want the possibility of locating normal components over a "wide" range of values and with a reasonable range of small and large variances (see Conley, Hansen, McCulloch, and Rossi (2008) for details on the prior assessment). Our default values are very diffuse.

$$\underline{a} = .016, \ \underline{c} = .17, \ \underline{\nu} = 2.004$$

We assess the prior on α to put prior mass on values of the number of unique components from 1 to at least 10 or more, though we note that the DP prior puts positive prior probability on up to N (the sample size) unique values or components.

In the Polya Urn method for drawing from the posterior distribution distribution in DP models, $\theta_i = (\mu_i, \Sigma_i)$ components are drawn for each observation. However, these values are clustered to a smaller number of unique values. The indicator variable can be formed from the set of draws of the errors distribution parameters and the set of unique values. This means that we can form a Gibbs sampler for the linear IV model with a DP process prior on the errors from the following steps:

$$\begin{gathered} \beta|\delta, ind, \{\theta_i\}, x, y, Z \\ \delta|\beta, ind, \{\theta_i\}, x, y, Z \\ \{\theta_i\}|\beta, \delta, x, y, Z \\ \alpha|I^* \end{gathered} \tag{5.14}$$

Given a set of draws of $\{\theta_i, i = 1, \ldots, N\}$, we can define the I^* unique values as $\left\{\theta_j^*, j = 1, \ldots, I^*\right\}$. The indicator vector, ind, is defined by $ind_i = j$ if $\theta_i = \theta_j^*$. The draws of β and δ in (5.14) are basically the same as for the normal model except that adjustments must be made for the means of the error terms and there are different means and variance terms depending on which unique value is associated with each observation. The *bayesm* routine, `rivDP`, implements the full Gibbs sampler including a so-called "remix" step that is not documented in (5.14).

Conley, Hansen, McCulloch, and Rossi (2008) consider the performance of the Bayesian IV procedure with DP errors under conditions of both normal and non-normal errors and weak and strong instruments. Performance is measured by the efficiency of a point estimator as well as the coverage of HPD intervals. There is virtually no loss to the use of the DP prior in the sense that, under normal errors, the inference on the structural parameters, β, is the same under DP or normal priors. However, under non-normal errors, the DP prior adapts to the non-normality and outperforms Bayes IV based on a single component normal prior. Comparisons to state-of-the-art classical methods for non-normal errors and weak instruments

show that the Bayes procedure extracts much more information from the sample.

5.2 Strategically Determined X Values in A Hierarchical Setting

In some marketing situations, marketing variables are customized at the cross-sectional unit level. Examples include customizing trade promotions or wholesales prices to specific markets, targeting and customization of coupons to specific consumers, and allocation of sales force differentially across accounts. In a typical hierarchical setting, we start with a conditional response model of the general form.

$$p\left(y_{ht}|x_{ht},\theta_h\right) \tag{5.15}$$

Implicitly, the standard analyses of this situation, consider the distribution of x_{ht} to be independent of θ_h. However, a more general approach would be to simultaneously model the sales response model and the determination of the marketing mix, taking into account that the distribution of x_{ht} may depend on θ_h.

$$\begin{array}{c} p\left(y_{ht}|x_{ht},\theta_h\right) \\ p\left(x_{ht}|\theta_h,\tau\right) \end{array} \tag{5.16}$$

This approach is a generalization of the models developed by Chamberlain (1980) and Chamberlain (1984) and applied in a marketing context by Bronnenberg and Mahajan (2001). Chamberlain considers situations in which the x variables are correlated with random intercepts in a variety of standard linear and logit/probit models. Our random effects apply to all of the response model parameters and we can handle non-standard and non-linear models. However, the basic results of Chamberlain regarding consistency of the conditional modeling approach apply. Unless T grows, any likelihood-based estimator for the conditional model will be inconsistent. The severity of this asymptotic bias will depend on model, data and T. For small T, these biases have been documented to be very large.

The general data-augmentation and Metropolis Hasting MCMC approach is ideally suited

to exploit the conditional structure of (5.16). That is, we can alternate between draws of $\theta_h|\tau$ (here we recognize that the $\{\theta_h\}$ are independent conditional on τ) and $\tau|\{\theta_h\}$. With some care in the choice of the proposal density, this MCMC approach can handle a very wide range of specific distributional models for both the conditional and marginal distributions.

To further specify the model in (5.16), it is useful to think about the interpretation of the parameters in the θ vector. We might postulate that in the marketing mix application, the important quantities are the level of sales given some "normal" settings of x (e.g. the baseline sales) and the derivative of sales wrt various marketing mix variables. In many situations, decision makers are setting marketing mix variables proportional to the baseline level of sales. More sophisticated decision makers might recognize that the effectiveness of the marketing mix is also important in allocation of marketing resources. This means that the specification of the marginal distribution of x should make the level of x a function of the baseline level of sales and the derivatives of sales with respect to the elements of x.

Manchanda, Rossi, and Chintagunta (2004) consider a special case of (5.16) in which the sales response model is a Negative Binomial Regression for the counts of prescriptions made by physicians. x includes the number of sales calls during the same period of time that prescriptions are monitored. The number of sales calls is modeled as strategically determined as a function of the sales call ("detail") responsiveness. This means that both the level and changes in the number of sales calls are informative regarding the effect of this variable. The additional information that is available from the modeling of the x variable as a function of θ_h is considerable and is a more important aspect of the problem than possible biases from the fact that x is dependent on the random coefficients (see Yang, Chen, and Allenby, 2003).

6 Conclusions

In this chapter, we have reviewed applications of Bayesian methods and models in marketing. Marketing applications highlight two under-emphasized aspects of the Bayesian paradigm. Due to the low information content and discreteness of disaggregate marketing data, informative priors are essential and require careful assessment. Flexibility in the specification of prior

distributions, particularly in the hierarchical setting is very important. Finally, marketing applications require that models of the consumer decision process be implemented. This gives rise to non-standard likelihoods. The simulation-based methods, now dominant in Bayesian work, free the investigator from reliance on standard models and priors. We regard marketing applications as a stimulating source of new models and a severe stress test for existing models and methods.

References

Allenby, G. M., M. J. Garratt, and P. E. Rossi (2010): "A Model for Trade-Up and Change in Considered Brands," *Marketing Science*, 29(1), 40–56.

Allenby, G. M., and P. E. Rossi (1991): "Quality Preceptions and Asymmetric Switching Between Brands," *Marketing Science*, 10(3), 185–204.

——— (1993): "A Bayesian Approach to Estimating Household Parameters," *Journal of Marketing Research*, 30(2), 171–182.

——— (1999): "Marketing Models of Consumer Heterogeneity," *Journal of Econometrics*, 89, 57–78.

Antoniak, C. E. (1974): "Mixtures of Dirichlet Processes with Applications to Bayesian Nonparametric Problems," *The Annals of Statistics*, 2(6), 1152–1174.

Avriel, M. (1976): *Nonlinear Programming: Analysis and Methods*. Prentice-Hall.

Bernardo, J. M., and A. F. M. Smith (1994): *Bayesian Theory*. John Wiley & Sons.

Berry, S., J. Levinsohn, and A. Pakes (1995): "Automobile Prices in Market Equilibrium," *Econometrica*, 63(4), 841–890.

Bhat, C. R. (2005): "A Multiple Discrete-Continuous Extreme Value Model: formulation and application to discretionary time-use decisions," *Transportation Research Part B*, 39, 679–707.

——— (2008): "The Multiple Discrete-continuous Extreme Value Model: role of utility function parameters, identification considerations, and model extensions," *Transportation Research Part B*, 42, 274–303.

Bronnenberg, B. J., and V. Mahajan (2001): "Multimarket Data: Joint Spatial Dependence in Market Shares and Promotional Variables," *Marketing Science*, 20(3), 284–299.

Chamberlain, G. (1980): "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 47, 225–238.

——— (1984): "Panel Data," in *Handbook of Econometrics*, ed. by Z. Griliches, and M. Intrilligator, vol. 2. North-Holland.

Chandukala, S. R., J. Kim, T. Otter, P. E. Rossi, and G. M. Allenby (2007): "Choice Models in Marketing: Economic Assumptions, Challenges and Trends," *Foundations and Trends in Marketing*, 2(2), 97–184.

Chen, Y., and S. Yang (2007): "Estimating Disaggregate Models Using Aggregate Data Through Augmentation of Individual Choice," *Journal of Marketing Research*, 44, 613–621.

Chib, S. (2010): "MCMC Methods," in *Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. V. Dijk. Oxford University Press.

Chib, S., and E. Greenberg (1998): "Analysis of Multivariate Probit Models," *Biometrika*, 85(2), 347–361.

Conley, T. G., C. B. Hansen, R. E. McCulloch, and P. E. Rossi (2008): "A Semi-Parametric Bayesian Apporach to the Instrumental Variable Problem," *Journal of Econometrics*, 144, 276–305.

Dubé, J.-P. (2004): "Multiple Discreteness and Product Differentiation: Demand for Carbonated Soft Driks," *Marketing Science*, 23(1), 66–81.

Edwards, Y., and G. M. Allenby (2003): "Multivariate Analysis of Multiple Response Data," *Journal of Marketing Research*, 40(3), 321–334.

Fruhwirth-Schnatter, S. (2006): *Finite Mixture and Markov Switching Models*. Springer.

Geweke, J. (1991): "Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints," in *Computing Science and Statistics: Proceedings of the 23rd Symposium*, ed. by E. M. Keramidas, pp. 571–78.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*. John Wiley & Sons.

Gilbride, T. J., and G. M. Allenby (2004): "A Choice Model with Conjunctive, Disjunctive, and Compensatory Screening Rules," *Marketing Science*, 23(3), 391–406.

Griffin, J., F. Quintana, and M. F. J. Steel (2010): "Flexible and Nonparametric Modelling," in *Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. V. Dijk. Oxford University Press.

Guadagni, P. M., and J. D. C. Little (1983): "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2(3), 203–238.

Hajivassiliou, V., D. L. McFadden, and P. Ruud (1996): "Simulation of Multivariate Normal Rectangle Probabilities and their Derivatives," *Journal of Econometrics*, 72, 85–134.

Hausman, J., and D. A. Wise (1978): "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preferences," *Econometrica*, 46(2), 403–426.

Imai, K., and D. A. van Dyk (2005): "A Bayesian Analysis of the Multinomial Probit Model using Marginal Data Augmentation," *Journal of Econometrics*, 124, 311–334.

Jiang, R., P. Manchanda, and P. E. Rossi (2009): "Bayesian Analysis of Random Coefficient Logit Models Using Aggregate Data," *Journal of Econometrics*, 149, 136–148.

Keane, M. P. (1994): "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62(1), 95–116.

Kim, J., G. M. Allenby, and P. E. Rossi (2002): "Modeling Consumer Demand for Variety," *Marketing Science*, 21(3), 229–250.

Koop, G. (2003): *Bayesian Econometrics*. John Wiley & Sons.

Lancaster, T. (2004): *An Introduction to Modern Bayesian Econometrics*. Blackwell.

Li, M., and J. L. Tobias (2010): "Bayesian Methods in Microeconometrics," in *Handbook of Bayesian Econometrics*, ed. by J. Geweke, G. Koop, and H. V. Dijk. Oxford University Press.

Manchanda, P., A. Ansari, and S. Gupta (1999): "The "Shopping Basket": A Model for Multicategory Purchase Incidence Decisions," *Marketing Science*, 18(2), 95–114.

Manchanda, P., P. E. Rossi, and P. K. Chintagunta (2004): "Response Modeling with Nonrandom Marketing-Mix Variables," *Journal of Marketing Research*, 41, 467–478.

McCulloch, R. E., N. G. Polson, and P. E. Rossi (2000): "A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters," *Journal of Econometrics*, 99, 173–193.

McCulloch, R. E., and P. E. Rossi (1994): "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 207–240.

McFadden, D. L. (1981): "Econometric Models of Probabilistic Choice," in *Structural Analysis of Discrete Choice*, ed. by M. Intrilligator, and Z. Griliches, pp. 1395–1457. North-Holland.

Morrison, D. G., and D. C. Schmittlein (1988): "Generalizing the NBD Model of Customer Purchases: What Are the Implications and Is It Worth The Effort?," *Journal of Business and Economic Statistics*, 6(2), 145–159.

Musalem, A., E. T. Bradlow, and J. S. Raju (2009): "Bayesian Estimation of Random-Coefficients Choice Models Using Aggregate Data," *Journal of Applied Econometrics*, 24, 490–516.

Orme, B. K. (2009): *Getting Started with Conjoint Analysis*. Research Publishers, LLC.

Pudney, S. E. (1989): *Modeling Individual Choice: The Econometrics of Corners, Kinks, and Holes*. Basil Blackwell.

R (2009): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing.

Roberts, G. O., and J. S. Rosenthal (2001): "Optimal Scaling for Various Metropolis-Hastings Algorithms," *Statistical Science*, 16(4), 351–367.

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

Rossi, P. E., Z. Gilula, and G. M. Allenby (2001): "Overcoming Scale Usage Heterogeneity: A Bayesian Hierarchical Approach," *Journal of the American Statistical Association*, 96(453), 20–31.

Rossi, P. E., and R. E. McCulloch (2008): *bayesm: Bayesian Inference for Marketing/Micro-Econometrics*2.2-3 edn.

Rossi, P. E., R. E. McCulloch, and G. M. Allenby (1996): "The Value of Purchase History Data in Target Marketing," *Marketing Science*, 15(4), 321–340.

Sonnier, G., A. Ainslie, and T. Otter (2007): "Heterogeneity Distributions of Willingness-to-Pay in Choice Models," *Quantitative Marketing and Economics*, 5, 313–331.

Train, K. E. (2003): *Discrete Choice Methods with Simulation*. Cambridge University Press.

Yang, S., Y. Chen, and G. M. Allenby (2003): "Bayesian Analysis of Simultaneous Demand and Supply," *Quantitative Marketing and Economics*, 1, 251–275.

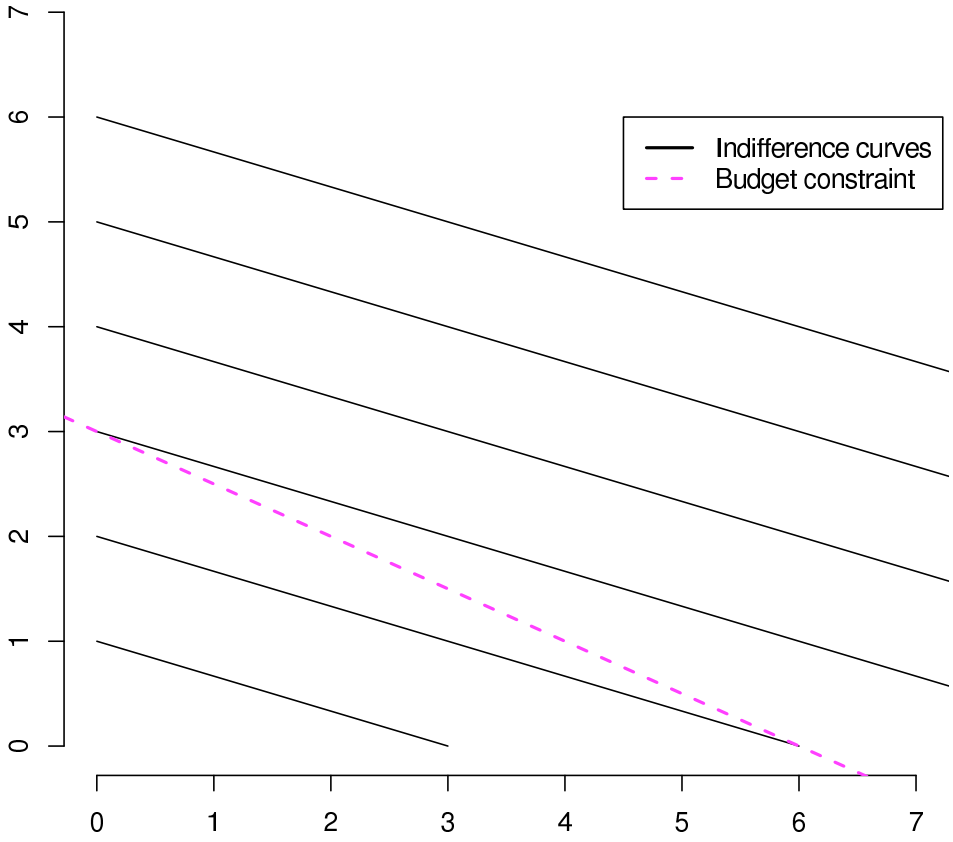

Figure 1: Homothetic Linear Utility

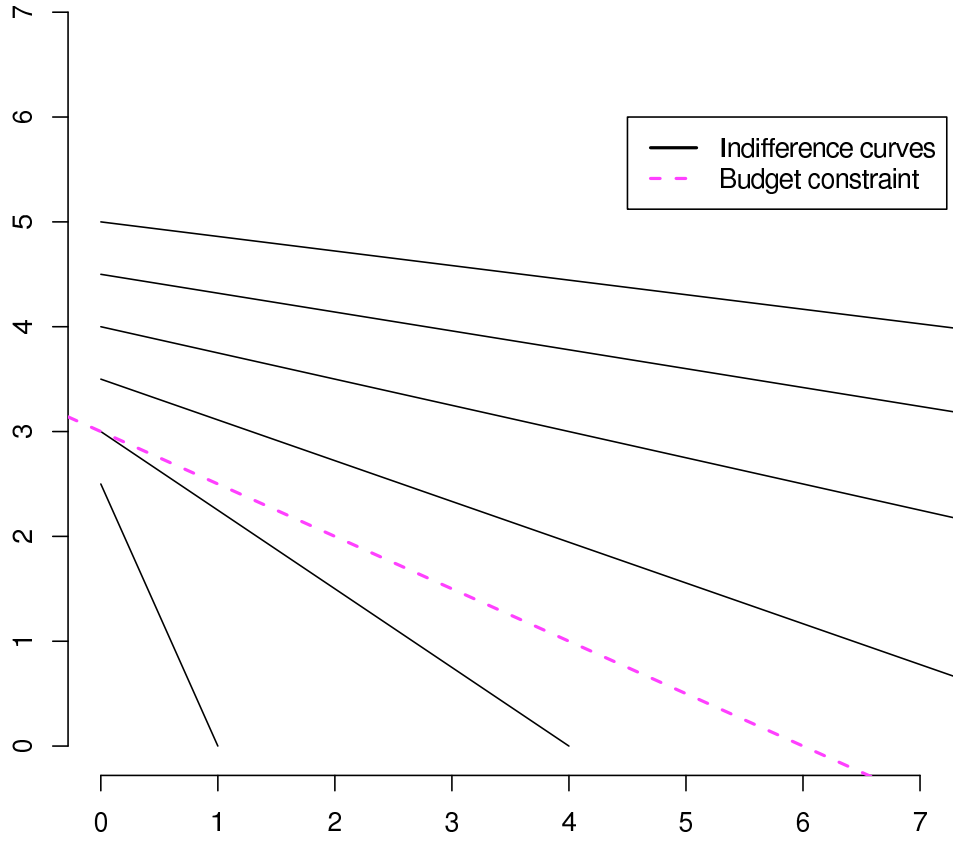

Figure 2: Non-homothetic Linear Utility

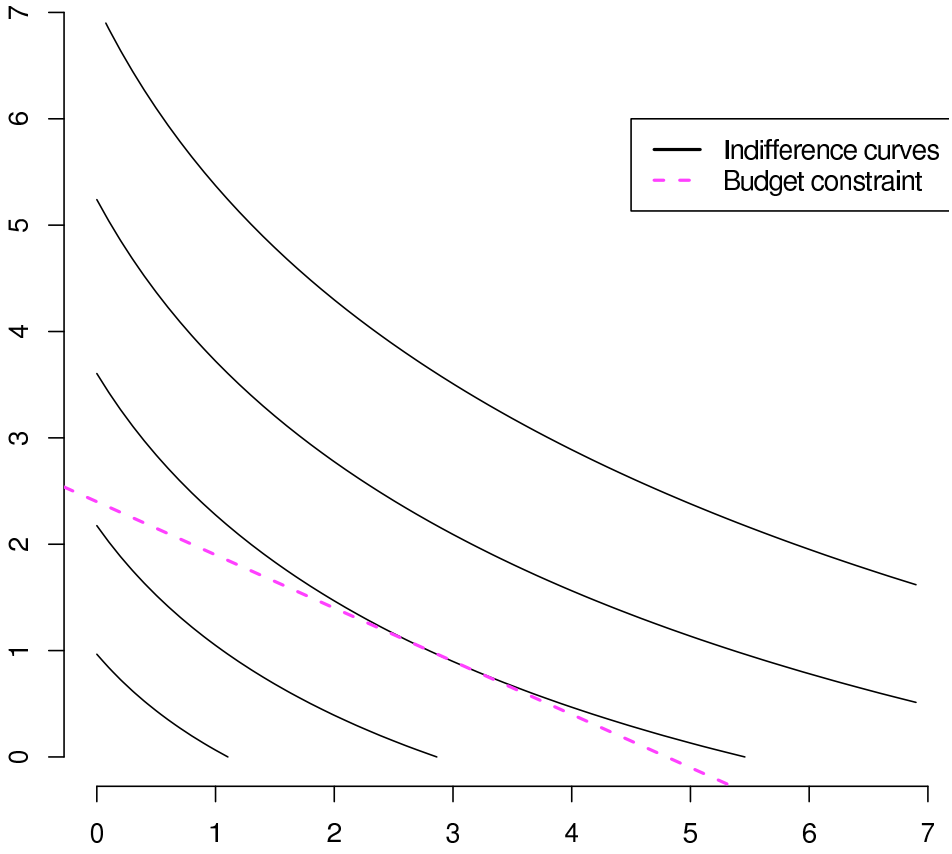

Figure 3: Non-linear Utility

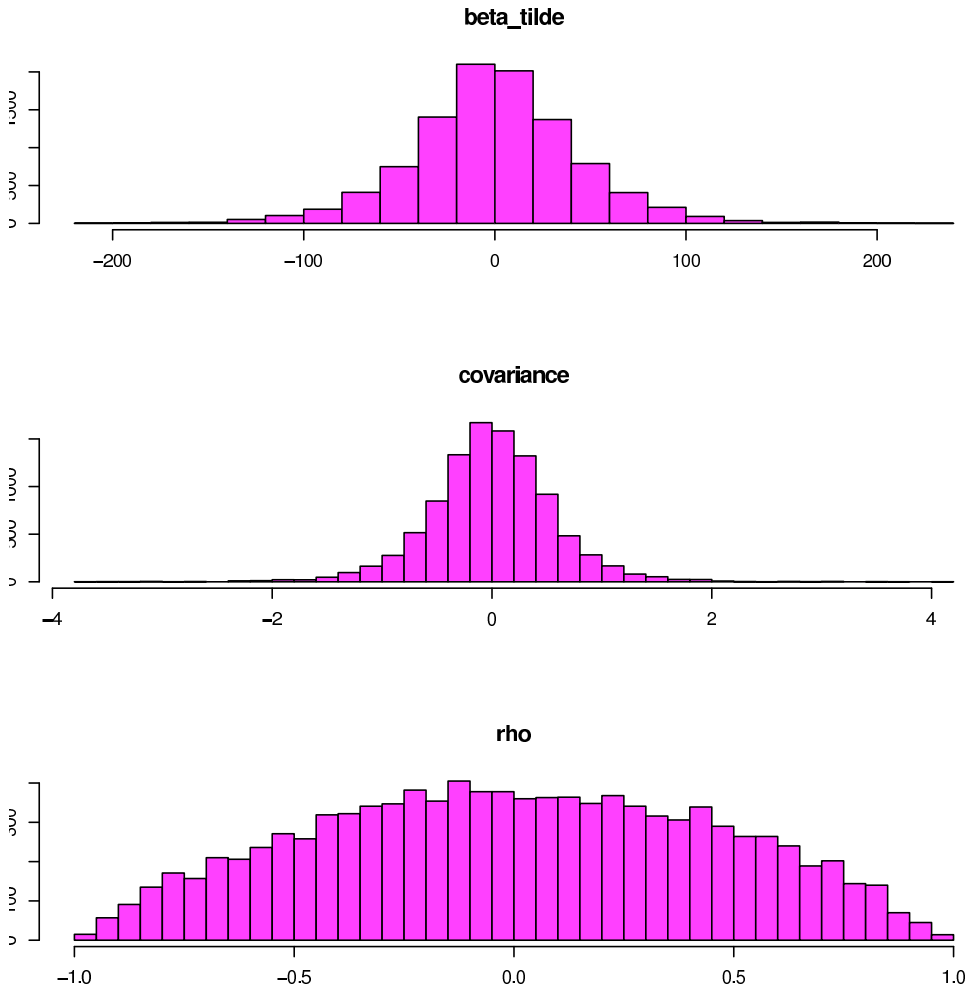

Figure 4: Default Prior for MNP Model