Chapter 4 Unit-Level Models and Discrete Demand

> *Using this Chapter*
>
> This chapter reviews models for discrete data. Much of the disaggregate data collected in marketing has discrete aspects to the quantities of goods purchased. Sections 4.1-4.3 review the latent variable approach to formulating models with discrete dependent variables, while section 4.4 derives models based on a formal theory of utility maximization. Those interested in Multinomial Probit or Multivariate Probit models should focus on sections 4.2 and 4.3. Section 4.2.1 provides material on understanding the difference between various Gibbs samplers proposed for these models and can be omitted by those seeking a more general appreciation. Section 4.4 forges a link between statistical and economic models and introduces demand models which can be used for more formal economic questions such as welfare analysis and policy simulation.

We define "unit-level" as the lowest level of aggregation available in a data set. For example, retail scanner data is available at many levels of aggregation. The researcher might only have regional or market level aggregate data. Standard regression models can suffice for this sort of highly aggregated data. However, as the level of aggregation declines to consumer level, sales response becomes more and more discrete. There are a larger number of zeroes in this data and often only a few integer-valued points of support. If, for example, we examine the prescribing behavior of a physician over a short period of time, this will be a count variable. Consumers often choose to purchase only a small number of items from a large set of alternatives. The goal of this chapter is to investigate models appropriate for disaggregate data. The common characteristic of these models will be the ability to attach lumps of probability to specific outcomes. It should also be emphasized, that even if the goal is to analyze only highly aggregate data, the researcher could properly view this data as arising from individual level decisions aggregated up to form the data observed. Thus, individual level demand models and models for the distribution of consumer preferences (the focus of chapter 5) are important even if the researcher only has access to aggregate data.

4.1 Latent Variable Models

We will take the point of view that discrete dependent variables arise via partial observation of an underlying continuous variable. For example, the binary probit model discussed in chapter 3 was formulated as a binary outcome which is simply the sign (0 if negative, 1 if positive) of the dependent variable in a normal regression model. There are three advantages of a latent variable formulation: 1. The latent variable formulation is very flexible and is capable of generating virtual any sort of outcome that has a discrete component; 2. The latent variable formulation allows for easy formulation of MCMC algorithms using data augmentation; and 3. In some situations, the latent variable can be given a Random Utility interpretation which relates latent variable models to the formal econometric specification of demand models based on utility maximization.

We can start with a normal regression model as the model for the latent variables.

$$z_i = X_i\delta + v_i \tag{4.1.1}$$

The outcome variable, y_i, is modeled as a function of the latent variable z_i.

$$y_i = f\left(z_i\right) \tag{4.1.2}$$

In order to create discrete y, the f function must be constant over some regions of the space in which z resides. We distinguish between the cases in which z is univariate and multivariate. For the case of univariate z, standard models such as the binary logit/probit and ordered probit models can be obtained. The binary probit model assumes $f(z) = I\ (z > 0)$ and v has a normal distribution. The binary logit assumes the same f function and extreme value type I with constant scale for v.

Ordered models can be obtained by taking the f function to be an indicator function of which of several intervals z resides in.

(4.1.3) $$f\left(z\right)=\sum_{c=1}^{C+1} c\times I\left(\gamma_{c-1}<z\leq\gamma_c\right)$$

y is multinomial with values 1, …, C and $\gamma_0=-\infty$ and $\gamma_{C+1}=\infty$. The ordinal nature of y is enforced by the assumption of a uni-dimensional latent variable. As $x_i^{'}\delta$ increases, the probability of obtaining higher values of y increases. By allowing for arbitrary cut-off points, the ordered model has flexibility to allow for different probabilities of each integer value of y conditional on x. For convenience in implementation of MCMC algorithms, we typically use an ordered probit model in which the latent variable is conditionally normal. As with the binary probit/logit models there is little difference between ordered probits and other models obtained by other distributional assumptions on v.

In other situations, y is multinomial but we do not wish to restrict the multinomial probabilities to be monotone in $X_i\delta$. In these cases, the underlying latent structure must be multivariate in order to provide sufficient flexibility to accommodate general patterns of the influence of the x variables on the probability that y takes on various values. That is, if we increase x, this may increase the probability of some value of y and decrease the probability of others in no particular pattern. Consider the multinomial case in which y can assume one of p values associated with the integers 1, …, p. In this case, we assume that z has a continuous, p dimensional distribution. A convenient choice of f is the indicator of the component with the maximum value of z.

(4.1.4) $$f\left(z\right)=\sum_{j=1}^{p} j\times I\left(\max\left(z\right)=z_j\right)$$

The multinomial distribution is appropriate for choice situations in which one of p alternatives is chosen. However, in some situations, multiple choices are a possible outcome. For example, we may observe consumers purchasing more than one variety of a product class on the same purchase occasion. In a survey context, a popular style of question is the "m choose k" or m alternatives are given and the questionnaire requests that the respondent "check all that apply." For these situations, a multivariate discrete choice model is appropriate. We represent this situation by allowing y to be p dimensional with each component having only two values. $y=\left(y_1,\ldots,y_p\right);\ y_i=0,1$. For this model, we must also use a multivariate latent variable and this becomes the direct generalization of the "univariate" binary model.

$$f:R^p \to R^p;\ f\left(z\right)=\begin{pmatrix} I\left(z_1>0\right) \\ \vdots \\ I\left(z_p>0\right) \end{pmatrix} \tag{4.1.5}$$

In some contexts, the response variable might reasonably be viewed as consisting of both discrete and continuous components. For example, we may observe both product choice and quantity. On many occasions, the consumer does not purchase the product but on some occasions a variable quantity is purchased. The quantity purchased is also discrete in the sense that most products are available in discrete units. However, we might assume that a continuous variable is a reasonable approximation to the quantity demanded conditional on positive demand. These situations can also be accommodated in the latent variable framework. For example, we can specify a simple Tobit model as

$$f\left(z\right)=z\times I\left(z>0\right) \tag{4.1.6}$$

However, we will defer a formal discussion of the mixed discrete continuous models until the introduction of a utility maximization model and a random utility framework.

Each of the models introduced in (4.1.1)-(4.1.5) has associated identification problems. These problems stem from the invariance of the $f\left(\bullet\right)$ functions to location or scale transformations of z, i.e. $f(cz)=f(z)$ or $f(z+k)=f(z)$. In the binary models, there is a scaling problem for the error, v. Changing the variance of v will not change likelihood as multiplication of z by a positive constant in (4.1.2) will not alter y. Typically, this is solved by imposing a scale restriction such as Var(v) =1. In the ordered model in (4.1.3), there is both a scaling and location problem. Even if we fix the scale of v, we can add a constant to all of the cut-offs and subtract the same constant from the intercept in δ and leave the likelihood unchanged. For this reason, investigators typically set the first cutoff to zero, $\gamma_1 = 0$. In the multinomial model(4.1.4), there is both a location and scale problem. If we add a constant to every component of z or if we scale the entire z vector by a positive constant, then the likelihood is unchanged. We will discuss identification of multinomial models in section 4. 2. For the multivariate model in (4.1.5), we recognize that we can scale each of the components of z by a *different* positive amount and leave the likelihood unchanged. This means that only the correlation matrix of z is identified.

4.2 MNP Model

If latent utility is conditionally normal and we observe outcome value i (of p) if max(z) = z_i, then we have specified the Multinomial Probit model.

$$\begin{aligned} & y_i = f(z_i) \\ (4.2.1) \qquad & f(z_i) = \sum_{j=1}^{p} j \times I\left(\max(z_i) = z_{ij}\right) \\ & z_i = X_i\delta + v_i \quad v_i \sim \text{iid}\, N(0,\Omega) \end{aligned}$$

If the multinomial outcome variable y represents consumer choice among p mutually exclusive alternatives, then X matrix consists of information about the attributes of each of the choice alternatives as well as covariates which represent the characteristics of the consumer making choices.

The general X would have the structure $X_i = \left[\left(1, d_i'\right) \otimes I_p,\ A_i\right]$ where d is a vector of "consumer" characteristics and A is a matrix of choice attributes. `createX` in our package, ***bayesm***, can be used to create the X array with this structure. For example, we might have information on the price of each alternative as well as the income of each consumer making choices. Obviously, we can only identify the effects of consumer covariates with a sample taken across consumers. In most marketing contexts, we obtain a sample of consumers in panel form. We will consider modeling differences between consumers in chapter 5. In a modern hierarchical approach, covariates are included in the distribution of parameters across consumers and not directly in the X matrix above. For expositional purposes, we will consider the case in which X contains only alternative specific information. In typical applications, δ contains alternative-specific intercepts as well as marketing mix variables for each alternative. It is common to assume that the coefficient on the marketing mix variables is the same across choice alternatives. The random utility derivation of multinomial choice models provides a rationale for this which we will develop in section 4.3. The linear model for the latent variable in (4.2.1) is a SUR model, written in a somewhat non standard form, and with restrictions that some of the regression coefficients are the same across regressions. For example, with a set of intercepts and one marketing mix variable, $X_i = \left[I_p \quad m_i\right]$, where I is a p x p identity matrix and m is a vector with p values of one marketing mix variable for each of p alternatives.

As indicated in section 4.1, the model in (4.2.1) is not identified. In particular, if we add a scalar random variable to each of the p latent regressions, the likelihood remains unchanged. That is, $\mathrm{Var}\left(z_i \middle| X_i, \delta\right) = \mathrm{Var}\left(z_i + u\iota \middle| X_i, \delta\right)$ where u is a scalar random variable. This is true only if Ω is unrestricted. If, for example, Ω is diagonal, then adding a scalar random variable will change the covariance structure by adding correlation between choice alternatives and the elements of the diagonal of Ω will be identified. In the case of unrestricted Ω, it is common practice to subtract the pth equation from each of the first (p-1) equations to obtain a differenced system. The differenced system can be written

$$\begin{aligned} & w_i = X_i^d \beta + \varepsilon_i \quad \varepsilon_i \sim N(0,\Sigma) \\ \text{(4.2.2)} \qquad & w_{ij} = z_{ij} - z_{ip};\ X_i = \begin{bmatrix} x_{i,1}' \\ \vdots \\ x_{i,p}' \end{bmatrix};\ X_i^d = \begin{bmatrix} x_{i,1}' - x_{i,p}' \\ \vdots \\ x_{i,p-1}' - x_{i,p}' \end{bmatrix};\ \varepsilon_{ij} = v_{ij} - v_{ip} \\ & y_i = f(w) = \sum_{j=1}^{p-1} j \times I\left(\max(w_i) = w_{ij}\right) + p \times I(w < 0) \end{aligned}$$

We also note that if δ contains intercepts, then the intercept corresponding to the pth choice alternative has been set to zero and β contains all of the other elements of δ.

The system in (4.2.2) is still not identified as it is possible to multiply w by a positive scalar and leave the likelihood of the observed data unchanged. In other words, $f(w) = f(cw)$. This identification problem is a normalization problem. We have to fix the scale of the latent variables in order to achieve identification. It is common to set $\sigma_{11} = 1$ to fix the scale of w. However, it is also possible to achieve identification by setting one of the components of the β vector to some fixed value. For example, if price were included as a covariate, we might fix the price coefficient at -1.0 Of course, this requires a prior knowledge of the sign of one of the elements of the β vector. In the case of models with price, we might feel justified in imposing this exact restriction on the sign of the price coefficient, but in other cases we might be reluctant to make such an imposition.

In classical approaches to the MNP, the model in (4.2.2) is reparameterized to the identified parameters $\left(\tilde{\beta}, \tilde{\Sigma}\right)$ by setting $\sigma_{11} = 1$. In a Bayesian approach, it is not necessary to impose identification restrictions. With proper priors, we can define the posterior in the unidentified space, construct an MCMC method to navigate this space and then simply "margin down" or report the posterior distribution of the identified quantities. In many cases, it is easier to define an MCMC method on the unrestricted space. Moreover, it turns out that the unrestricted sampler will often have better mixing properties.

In 3.7, we saw how data augmentation can be used to propose a Gibbs sampler for the binary probit model. These same ideas can be applied to the MNP model. If β, Σ are a priori independent, the DAG for the MNP model is given by

(4.2.3)
$$\begin{tikzcd}
\Sigma \arrow[dr] & & \\
& w \arrow[r] & y \\
\beta \arrow[ur] & &
\end{tikzcd}$$

McCulloch and Rossi (1994) propose a Gibbs sampler based on cycling through three conditional distributions.

(4.2.4)
$$\begin{aligned}
&w \mid \beta, \Sigma, y, X_i^d \\
&\beta \mid \Sigma, w \\
&\Sigma \mid \beta, w
\end{aligned}$$

This sampler produces a chain with $w, \beta, \Sigma \mid y$ as its stationary distribution. We can simply marginalize out w if the posterior of β, Σ is desired. We note that w is sufficient for β, Σ and this is why it is not necessary to include y as a conditioning argument in the second and third distributions in (4.2.4). Since the underlying latent structure is a normal Multivariate regression model, we can use the standard theory developed in chapter 2 to draw β and Σ. The only difficulty is the draw of w. Given β, Σ, the $\{w_i\}$ are independent, p-1 dimensional truncated normal random vectors. The regions of truncation are the R_{y_i}, as defined in

(4.2.11). Direct draws from truncated multivariate normal random vectors are difficult to accomplish efficiently. The insight of McCulloch and Rossi (1994) is to recognize that one can define a Gibbs sampler by breaking each draw of w_i into a sequence of p-1 univariate truncated normal draws by cycling through the w vector.

That is, we do not draw $w_i \left| \beta, \Sigma \right.$ directly. Instead, we draw from each of the (p-1) truncated univariate normal distributions.

(4.2.5)
$$\begin{aligned}
& w_{ij} \left| w_{i,-j}, y_i, \beta, \Sigma \sim N\left(m_{ij}, \tau_{jj}^2\right) \times \right. \\
& \left[I\left(j = y_i\right) I\left(w_{ij} > \max\left(w_{i,-j}, 0\right)\right) + I\left(j \neq y_i\right) I\left(w_{ij} < \max\left(w_{i,-j}, 0\right)\right)\right] \\
& m_{ij} = x_{ij}^{d'}\beta + F'\left(w_{i,-j} - X_{i,-j}^{d}\beta\right) \\
& F = -\sigma^{j,j}\gamma_{j,-j} \\
& \tau_{j,j}^2 = 1/\sigma^{j,j}
\end{aligned}$$

$\sigma^{i,j}$ denotes the (i, j) th element of Σ^{-1} and

(4.2.6)
$$\Sigma^{-1} = \begin{bmatrix} \gamma_1' \\ \vdots \\ \gamma_{p-1}' \end{bmatrix}$$

The univariate distribution in (4.2.5) is a truncated normal distribution. $\gamma_{j,-j}$ refers to the jth row of Σ^{-1} with the jth element deleted. $X_{i,-j}^{d}$ is the matrix, X_i, with the jth column deleted. We start with the first element (j=1) and "Gibbs thru" each observation, replacing elements of w, one by one, until the entire vector is updated. In *bayesm*, this is done in C code. To implement this sort of sampler in R, `condMom` can be useful to compute the right moments for each univariate truncated normal draw.

To implement full Gibbs sampler for the MNP model, we need to specify a prior over the model parameters, β, Σ. The model parameters are not identified without further restrictions due to the scaling problem for the MNP model. The identified parameters are obtained by normalizing with respect to one of the diagonal elements of Σ.

(4.2.7) $$\tilde{\beta} = \beta / \sqrt{\sigma_{11}}; \; \tilde{\Sigma} = \Sigma / \sigma_{11}$$

For ease of interpretation, we will report the correlation matrix and vector of relative variances as the set of identified covariance parameters. One approach would put a prior on the full set of un-identified parameters as in McCulloch and Rossi (1994).

(4.2.8) $$\beta \sim N\left(\bar{\beta}, A^{-1}\right) \;\; \Sigma \sim IW\left(\upsilon, V_0\right)$$

(4.2.8) induces a prior on the identified parameters in (4.2.7). The induced prior on the identified parameters is not in standard form.[1] Imai and van Dyk (2004) suggest a very similar prior but with the advantage that the prior induced on the identified regression coefficients is more easily interpretable.

(4.2.9) $$\beta | \Sigma \sim N\left(\sqrt{\sigma_{11}}\,\bar{\beta}, \sigma_{11} A^{-1}\right); \;\; \Sigma \sim IW\left(\upsilon, V_0\right)$$

With the prior in (4.2.9), $\tilde{\beta} \sim N\left(\bar{\beta}, A^{-1}\right)$. This prior specification makes assessment of the prior more straightforward. It also allows for the use of an improper prior on the identified regression coefficients.

Another approach would be to put a prior directly on the set of identified parameters and define a Gibbs sampler on the space $\left(w, \tilde{\beta}, \tilde{\Sigma}\right)$ as in McCulloch et al (2000). This

[1] McCulloch et al (2000) derive these distributions. Assessment of these priors for all but the highly diffuse case may be difficult. Under the prior in (4.2.8), $\tilde{\beta}$ and $\tilde{\Sigma}$ are not independent. In addition, the prior on $\tilde{\beta}$ can be skewed for non-zero values of $\bar{\beta}$.

requires a prior on the set of covariance matrices with 1,1 element set equal to 1. This is done via a reparameterization of Σ.

(4.2.10)
$$\Sigma = \begin{bmatrix} \sigma_{11} & \gamma' \\ \gamma & \Phi + \gamma\gamma' \end{bmatrix}$$

This reparameterization is suggested by considering the conditional distribution of $\varepsilon_{-1} \middle| \varepsilon_1 \sim N\left(\gamma\varepsilon_1, \Phi\right)$. This is a multivariate regression of ε_{-1} on ε_1 and the standard conjugate prior can be used for γ, Φ (as in section 2.8.5). σ_{11} can be set to 1 in order to achieve identification. This creates a prior over covariance matrices with the 1,1 element fixed at 1. However, the Gibbs sampler suggested by McCulloch et al (2000) (hereafter termed the ID MNP sampler) navigates the identified parameter space more slowly than the sampler of McCulloch and Rossi (1994). For this reason, we cover the approach of McCulloch and Rossi (1994) here.

Given the prior in (4.2.8), we can define the McCulloch and Rossi (1994) sampler, which we term the NID, for non-identified sampler, as follows

NID MNP GIBBS SAMPLER

Start with initial values, w_0, β_0, Σ_0

Draw $w_1 | \beta_0, \Sigma_0$ using (4.2.5)

Draw $\beta_1 \left| w_1, \Sigma_0 \sim N\left(\tilde{\beta}, V\right)\right.$

$$V = \left(X^{d*'}X^{d*} + A\right)^{-1} \quad \tilde{\beta} = V\left(X^{d*'}w^* + A\bar{\beta}\right)$$

$$\Sigma_0^{-1} = C'C$$

$$X_i^{d*} = C'X_i^d \quad w_i^* = C'w_i$$

$$X^d = \begin{bmatrix} X_1^d \\ \vdots \\ X_n^d \end{bmatrix}$$

Draw $\Sigma_1 \left| w_1, \beta_1 \right.$ using $\Sigma^{-1} \left| w, \beta \sim W\left(\upsilon + n, \left(V_0 + S\right)^{-1}\right)\right.$

$$S = \sum_{i=1}^{n} \varepsilon_i \varepsilon_i^{'}$$

$$\varepsilon_i = w_i - X_i^d \beta$$

Repeat as necessary

This sampler is implemented in `rmnpGibbs` in our package, *`bayesm`*.

To illustrate the functioning of the NID sampler, consider a simulated example from McCulloch and Rossi (1994).

N=1600, p=6. X~iidUnif(-2,2). $\beta = 2$. $\Sigma = \text{diag}\left(\sigma\right)\left(\rho\iota\iota' + \left(1-\rho\right) I_{p-1}\right)\text{diag}\left(\sigma\right)$. $\rho = .5$ and $\sigma' = \left(1,2,3,4,5\right)^{.5}$.

Figure 4.1 shows the MCMC trace plot and acfs for $\tilde{\beta}$ and ρ_{12}. These are achieved by normalizing the draws of the unidentified parameters appropriately. For example, to construct an MCMC estimate of the posterior distribution of $\tilde{\beta}$, we simply post-process the draws of β by dividing by the (1,1) element of Σ. Figure 4.1 shows the trace plots for the "thinned" draw sequence obtained by extracting every 100th draw from a draw sequence of 100, 000. However, the ACFs are computed using every draw. β was started at 0 and Σ at the identity matrix. Very diffuse priors were used, β ~ $N\left(0,100\right)$ and Σ ~ $IW\left(\nu = \left(p-1\right) + 2, \upsilon I_{p-1}\right)$. The MCMC sampler dissipates the initial conditions very rapidly but exhibits very high autocorrelation. One way to gauge the extent of this auto correlation is to compute the f factor or relative numerical efficiency which is the ratio of the variance of the mean of the MCMC draws to the variance assuming an iid sample. For $\tilde{\beta}$, f= 110 and, for ρ_{12}, f = 130. This means that our "sample" of 100,000 has the information content of approximately 9000 $\left(\approx {}^{100,000}\!/\!_{\sqrt{f}}\right)$ iid draws from the posterior. The posterior mean of $\tilde{\beta}$ is estimated to be 2.14 with a numerical standard error of .0084 and the posterior mean of ρ_{12} is .39 with a numerical standard error of .005. These numerical standard errors must be viewed relative to the posterior standard deviations of these quantities. If the posterior standard deviation is large, then it means that we can't make precise inferences about these parameters and we may be willing to tolerate larger numerical standard errors. In this example, the estimated posterior standard deviations for both $\tilde{\beta}$ and ρ_{12} are about 30 times the size of the numerical standard errors.

4.2.1 Understanding the Autocorrelation Properties of the MNP Gibbs Sampler

MNP Gibbs sampler exhibits higher autocorrelation than MCMC examples considered in chapter 3. As with all Gibbs samplers, high autocorrelation is created by dependence among the random variables being sampled. It appears that the introduction of the latent variables via data augmentation has created the problem. There must be high dependence between w and β and/or Σ. To verify and explore this intuition, consider the binary probit model which is a special case of the MNP model with p=2. For this case, we can easily identify two samplers: 1. the NID sampler[2] and 2. the ID sampler of Albert and Chib (see also chapter 3). Both samplers include both w and β, but the Albert and Chib sampler sets Σ=1 and navigates in the identified parameter space.

It is easier to develop an intuition for the high dependence between w and the model parameters in the case of the ID sampler as there is one fewer model parameter. Consider a simple case where there is only one X variable and this variable takes on only two values

[2] For p=2, our NID sampler is identical to that proposed by Van Dyk and Meng (2001).

(0,1). In this case, all observations fall into a 2 x 2 contingency table defined by X and Y. The latent variables are drawn from four different truncated distributions corresponding to cells of this table.

	Y=0	Y=1
X=0	$w_{0,0}$	$w_{0,1}$
X=1	$w_{1,0}$	$w_{1,1}$

To see why β and w can be very highly autocorrelated consider the case depicted in figure 4.2. If X=0, then w will be sampled from the normal distribution centered at zero which is depicted by the dotted density. For X=0 and Y=1, w will be sampled from the positive half normal and, for X=0 and Y=0, from the negative half normal. If β changes, these draws of w will remain unchanged in distribution. However, consider what happens to the draws of w for X=1. These draws are made from truncations of the normal distributions depicted by solid densities in figure 4.2. If β increases, then the draws of w will come from a normal distribution centered at a higher mean and truncated either from above by 0 or below by 0. These draws will be larger in expectation. This sets up a situation in which if β increases, w draws tend to increase for X=1 and remain unchanged for X=0. In turn, the posterior mean of β will increase as the least squares estimate of $\beta \mid w$ will simply be the difference in mean w for X=1 vs. X=0. The only force working against the strong dependence of w and β is the variability in the distribution of $w \middle| X = 1, \beta$.

If $\operatorname{Var}(x'\beta)$ is large relative to $\operatorname{Var}(\varepsilon)$, then the dependence between β and w will be high. In the ID sampler, $\operatorname{Var}(\varepsilon)$ is fixed at 1. However, in the NID sampler the variance of ε is drawn as a parameter. Large values of $\operatorname{Var}(\varepsilon)$ will reduce the dependence between w and β. The NID sampler can draw large values of $\mathrm{Var}(\varepsilon)$, particularly for very diffuse priors. Van Dyk and Meng (2001) observe that it would be optimal to set an improper prior for Σ in terms of the properties of the NID sampler. However, both the NID and ID samplers can exhibit high dependence between draws. Both samplers will perform poorly when $\mathrm{Var}(x'\beta)$ is large. This may be disconcerting to the reader – when the X variables are highly predictive of choice, the performance of both samplers degrades.

To illustrate the dependence between β and w, we consider some simple simulated examples. X is one column which is created by taking draws from a Bernoulli(.5) distribution. N = 100, Σ = 1 . We will simulate data corresponding to an increasing sequence of β values. The intuition is that, as β increases, we should observe an increase in dependence between w_i draws for X=1, Y=1 and β. Figure 4.3 plots the sample mean of the w draws for X=1, Y=1 against the β draw for the same ID MCMC iteration. Four simulated data sets with β = (1, 2, 3) were created. This correlation exceeds .99 for β = 3. Figure 4.4 provides the acfs for the draw sequence of β for each of the simulated datasets. The dependence between w and β results in extremely high autocorrelation for large values of β. It should be pointed out that a β value of 3 or larger is fairly extreme as this implies that the Prob(Y=1 | X =1) > .99[3]

However, the NID sampler performs much better as illustrated in figure 4.4. The top two ACFs show the ID and NID samplers for data simulated with β=2. The NID sampler exhibits much lower autocorrelation. For data simulated with β=3, the autocorrelation for the ID sampler becomes very extreme. In that case, the ID sampler output is some 15 times less informative than an iid sample. The NID sampler is highly autocorrelated but with output approximately twice as informative as the ID sampler. These results for the binary probit model suggest that the ID sampler for the MNP model proposed by McCulloch et al (2000) can exhibit very high autocorrelation. Nobile (2000) and Imai and Van Dyk (2004) point this out. The intuition developed here is that navigating in the unidentified parameter space can result in a chain with better mixing properties in the space of identified parameters. Nobile (1998) proposes a modification of the McCulloch and Rossi (1994) algorithm which includes a Metropolis step that increases the variability of the σ draws. Nobile observes improvement over the McCulloch and Rossi NID sampler under extreme conditions. As Van Dyk and Meng (2001) point out, any latent variable model that margins out to the right likelihood can be used as the basis for a data augmentation MCMC method[4]. It is entirely possible that even better performing chains can be constructed using different data augmentation schemes.

4.2.1 The likelihood for the MNP Model

The likelihood for the MNP model is simply the product of the relevant probabilities of the outcomes. Each of these probabilities involves integrating a normal distribution over a p-1 dimensional cone

$$\begin{aligned}
&\ell(\beta,\Sigma) = \prod_{i=1}^{n} \Pr\left(y_i \middle| X_i^d,\beta,\Sigma\right) \\
&\Pr\left(y_i \middle| X_i^d,\beta,\Sigma\right) = \int_{R_{y_i}} \varphi\left(w \middle| X_i^d\beta,\Sigma\right) dw \qquad (4.2.11)\\
&R_{y_i} = \begin{cases} \left\{w : w_{y_i} > \max\left(w_{-y_i},0\right)\right\} & \text{if } y_i < p \\ \left\{w : w < 0\right\} & \text{if } y_i = p \end{cases}
\end{aligned}$$

[3] It should be noted that Prob(Y=0 | X=0) = .5. This means that we are not in the situation in which the observations can be separated and the posterior with flat priors becomes improper.

w_{-j} denotes all elements in the w vector except the jth element (a (p-2) dimensional vector) and $\varphi(\bullet)$ is the multivariate normal density function. Figure 4.5 shows this situation for a 3 choice example. The contours correspond to a bivariate normal distribution centered at a mean determined by the X values for the ith observation. Three regions are shown on the figure, corresponding to each of the three choice alternatives. If y_i =1 then we must have $w_1 > \max\left(w_{2,}0\right)$. If y_i =2, then $w_2 > \max\left(w_{1,}0\right)$. If y_i =3, then $w_1, w_2 < 0$. Direct evaluation of the likelihood is only computationally feasible if these integrals can be evaluated accurately. The GHK method discussed in chapter 2 can be used to evaluate the likelihood in (4.2.11). The GHK method is designed to compute integral of a normal distribution over a rectangle defined by vectors of lower and upper truncation points. In order to use this method, we have to re-express the region defined in (4.2.11) as a rectangular region. To do so, we must define a matrix, specific to each choice alternative, that can be used to transform the w vector to a rectangular region.

For j= 1, …, p-1, define[5]

$$A_j = \begin{bmatrix} -I_{j-1} & \begin{matrix} 1 \\ \vdots \\ 1 \end{matrix} & 0 \\ 0 & & -I_{(p-1)-j} \end{bmatrix}$$

The condition $w_j > \max\left(w_{-j}\right)$ and $w_j > 0$ is equivalent to $A_j w > 0$. We can then re-express the inequalities by applying A_j to both sides of the latent variable equation

$$\begin{aligned} &A_j w = A_j \mu + A_j \varepsilon > 0 \\ &A_j \varepsilon > -A_j \mu \\ &\text{where } \mu = X^d \beta \end{aligned}$$

[4] In the sense that $p\left(y|\beta,\Sigma\right) = \int p\left(y,w|\beta,\Sigma\right)dw = \int p\left(y|w\right)p\left(w|\beta,\Sigma\right)dw$ for many latent models.

If we define $u = A_j \varepsilon$, then we can express choice probability as a multivariate normal integral truncated from below.

(4.2.12)
$$\begin{gathered} u > -A_j \mu \quad u \sim N\left(0, A_j \Sigma A_j^{'}\right) \\ \Pr\left(y = j \middle| X, \beta, \Sigma\right) = \int_{A_j X\beta}^{\infty} \varphi\left(u \middle| 0, A_j \Sigma A_j^{'}\right) du \end{gathered}$$

For the case of j=p, the region is already defined in terms of ε

(4.2.13)
$$\begin{gathered} \varepsilon < -\mu \\ \Pr\left(y = p \middle| X, \beta, \Sigma\right) = \int_{-\infty}^{-X\beta} \varphi\left(u \middle| 0, \Sigma\right) du \end{gathered}$$

The function, `llmnp`, in ***bayesm***, implements this approach.

We could use the likelihood evaluated via GHK to implement a Metropolis Chain for the MNP without data augmentation. However, this would require a good proposal function for both β, Σ. Experience with Metropolis methods for covariance matrices suggests that these methods are only useful for small dimensional problems.

4.3 Multivariate Probit Model

The Multivariate Probit model is specified by assuming the same multivariate regression model as for the MNP model but with a different censoring mechanism. We observe the sign of the components of the underlying p dimensional multivariate regression model.

(4.3.1)
$$\begin{gathered} w_i = X_i \beta + \varepsilon_i \qquad \varepsilon_i \sim N\left(0, \Sigma\right) \\ y_{ij} = \begin{cases} 1 \text{ if } w_{ij} > 0 \\ 0 \text{ otherwise} \end{cases} \end{gathered}$$

Here choice alternatives are not mutually exclusive as in the MNP model. The multivariate probit model has been applied to purchase of products in two different categories (Manchanda et al (1999)) or to surveys with pick j of p questions (Edwards and

[5] Note that if j=1 or j=p-1, then we simply omit the requisite I matrix (i.e. I_0 means nothing)!

Allenby(2003)). In the econometrics literature, the multivariate probit has been applied to a binary phenomenon which is observed over adjacent time periods (e.g. labor force participation observed for individual workers).

The identification problem in the Multivariate Probit can be different from the identification problem for the MNP depending on the structure of the X array. Consider the general case which includes intercepts for each of the p choice alternatives and covariates that are allowed to have different coefficients for each p choices.

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

z is a d x 1 vector of observations on covariates. Thus, X is a p x k matrix with k= p x d.

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

β_i , i = 1, .., d are p dimensional coefficient vectors. The identification problem arises from the fact that we can scale each of the p means for w with a different scaling constant without changing the observed data. This implies that only the correlation matrix of Σ is identified and that transformation from the unidentified to the identified parameters $\left((\beta,\Sigma) \rightarrow \left(\tilde{\beta}, R\right)\right)$ is defined by

(4.3.4) $$\begin{aligned} \tilde{B} &= \Lambda B \\ \tilde{\beta} &= \operatorname{vec}\left(\tilde{B}\right) \\ R &= \Lambda \Sigma \Lambda \end{aligned}$$

where

$$B = [\beta_1, \ldots, \beta_d]$$

$$\Lambda = \begin{bmatrix} 1/\sqrt{\sigma_{11}} & & \\ & \ddots & \\ & & 1/\sqrt{\sigma_{pp}} \end{bmatrix}$$

However, if the coefficients on a given covariate are restricted to be equal across all p choices, then there are fewer unidentified parameters. Then we cannot scale each equation by a *different* positive constant. This brings us back in to the same situation as in the MNP model where we must normalize by one of the diagonal elements of Σ. For example, we might have an attribute like price of the p alternatives under consideration. We might want to restrict the price attribute to have the same effect on w for each alternative. This amounts to the restriction that $\beta_{j1} = \beta_{j2} = \ldots \beta_{jp}$ for covariate j.

To construct a MCMC algorithm for the multivariate probit model, we can use data augmentation just as in the MNP model by adding w to the parameter space. We simply "Gibbs thru" the w vector using the appropriate conditional univariate normal distribution but with an upper (lower) truncation of 0 depending on the value of y.

$$\begin{aligned} & w_{ij} \,\middle|\, w_{i,-j}, y_i, \beta, \Sigma \sim N\left(m_{ij}, \tau_{jj}^2\right) \times \\ & \left[I\left(y_{ij} = 1\right) I\left(w_{ij} > 0\right) + I\left(y_{ij} = 0\right) I\left(w_{ij} < 0\right) \right] \\ & m_{ij} = x_{ij}^{'}\beta + F'\left(w_{i,-j} - X_{i,-j}\beta\right) \\ & F = -\sigma^{j,j}\gamma_{j,-j} \\ & \tau_{j,j}^2 = 1/\sigma^{j,j} \end{aligned} \tag{4.3.5}$$

Here the vector y is an np vector of indicator variables.

We must make a choice of whether to navigate in the unidentified (β, Σ) space or the identified $(\tilde{\beta}, R)$ space. The unidentified parameter space is larger by p dimensions than the identified space. The intuition developed from the MNP model and generalized by Van Dyk and Meng is that navigating in the higher dimensional unidentified parameter space with diffuse priors will produce a chain with superior mixing properties to a chain defined on the identified space. An additional complication will be the method for drawing valid R matrices. The algorithm of Chib and Greenberg (1998) or Barnard et al (2000) can be used to draw R. However, given the additional complication and computational cost of these methods and the fact that we expect these chains to have inferior mixing properties, we recommend using a more straightforward Gibbs sampler on the unidentified parameter space (see Edwards and Allenby (2003) for details).

NID MULTIVARIATE PROBIT GIBBS SAMPLER

Start with initial values, w_0, β_0, Σ_0

Draw $w^1 | \beta_0, \Sigma_0, y$ using (4.3.5)

Draw $\beta_1 \big| w_1, \Sigma_0 \sim N(\tilde{\beta}, V)$

$$V = \left(X^{d*\prime} X^{d*} + A\right)^{-1} \quad \tilde{\beta} = V\left(X^{d*\prime} w^* + A\bar{\beta}\right)$$

$$\Sigma_0^{-1} = C'C$$

$$X_i^{d*} = C' X_i^d \quad w_i^* = C' w_i$$

$$X^d = \begin{bmatrix} X_1^d \\ \vdots \\ X_n^d \end{bmatrix}$$

Draw $\Sigma_1 \left| w_1, \beta_1 \right.$ using $\Sigma^{-1} \left| w, \beta \sim W\left(\upsilon + n, \left(V_0 + S\right)^{-1}\right)\right.$

$$S = \sum_{i=1}^{n} \varepsilon_i \varepsilon_i^{'}$$

$$\varepsilon_i = w_i - X_i^d \beta$$

Repeat as necessary

This algorithm has been implemented in the function, `rmvpGibbs`, in *`bayesm`*.

To illustrate this sampler, we consider a data example from Edwards and Allenby (2003). This dataset is available in *`bayesm`* and can be loaded using the R command, `data(Scotch)`. 2218 respondents were given a list of 21 scotch brands and asked to indicate whether or not they drink these brands on a regular basis. The interest in this example is in understanding the correlation structure underlying brand choice. Correlation in the latent variable can be viewed as a measure of similarity of two brands. In this example, $X_i = I_{21}$ so that the β vector is simply a vector of intercepts or means for the latent variable w.

Figure 4.6 shows MCMC traces (every 20th draw from sequence of 20,000) and ACFs for two of the elements of β corresponding to popular blended whiskeys (note: the brand chosen most often is Chivas Regal so all other brands have a smaller intercepts). These plots show quick dissipation of the initial condition (β_0=0 and Σ_0=I). The autocorrelations of the intercept draws is small with a numerical efficiency roughly one half of that of an iid sample. Figure 4.7 shows MCMC traces (again every 20th draw from a sequence of 20,000) and ACFS for two of the correlations. These correlations are between two single malts and between the most popular blended whiskey (Chivas Regal) and a single malt. The single malts exhibit high correlation showing similarity in product branding and taste. There is a negative correlation between the single malt and the blended whiskey showing some divergence in consumer preference. The autocorrelations for the correlation parameters are higher than the intercept or mean parameters, but still less than for some of the MNP examples considered in section 4.2 (numerical efficiency here is 1/9 th of a random sample as compared to 1/11 th for the MNP examples). The intuition for the better performance of the NID sampler for the Multivariate Probit (in comparison to the NID MNP sampler) is that there is a higher dimensional unidentified parameter space improving the mixing characteristics of the sampler in the identified parameter space.

In our analysis of the scotch data, we used a relatively diffuse but proper prior: $\beta\sim$ $N\left(0,100I_p\right)$ and Σ ~ $IW\left(p+2,\left(p+2\right)I_p\right)$. Our analysis of the binary probit model developed an intuition that the diffusion of the prior on Σ would affect the performance of the sampler. In particular, as emphasized by Meng and Van Dyk (2001), mixing should be maximized by allowing the prior on Σ to be improper. However, as McCulloch et al (2000) have pointed out, improper priors on Σ can be extraordinarily informative on functions of Σ such as correlations. The implied marginal prior on each correlation will be U-shaped with substantial mass near -1 and 1.

We reran the analysis with an improper prior on Σ and show the results in figure 4.8. As expected, the improper prior produces better mixing, reducing the autocorrelations substantially. Numerical efficiency is now at 1/7 th of an iid sample. However, the improper prior on Σ is extraordinarily informative on the correlations changing both the location and tightness of the posterior distributions of the correlations. Figure 4.9 shows the prior distribution of a correlation for the barely proper case, $\Sigma \sim IW\left(p,pI_p\right)$. This "u-shaped" distribution puts high mass near high positive or negative correlations. The improper prior can be viewed as the limit of proper priors as the diffusion increases and, therefore, will be even more informative than the barely proper space, putting high mass on extreme values of correlation.

Thus, improper priors are very dangerous for this analysis even though they might receive attention due to mixing considerations. The problem here is that there is a conflict between prior settings which promote maximum mixing and substantive informative considerations. It would be desirable to produce a MCMC sampler with a prior that separates "tuning" from substantive informativeness.

4.4 Demand Theory and Models Involving Discrete Choice

The multinomial and multivariate choice models developed in section 4.2 and 4.3 are simply statistical models that use a latent variable device to link discrete outcomes to covariate information. Since the outcomes in many marketing situations are the result of customer decisions, we should relate these statistical models to models of optimizing behavior and consumer demand theory. If we assume that consumers are maximizing utility subject to a budget constraint, then our modeling efforts focus on choice of the utility function and assumptions regarding the distribution of unobservables.

In much of standard demand theory (c.f. Deaton and Muellbauer (1980)), utility functions are chosen so that only interior solutions are obtained. That is, consumers faced with a set of goods would purchase all goods in various amounts. For aggregate data or for highly aggregated goods (such as food and housing), these utility functional forms are appropriate. However, at the level of the individual customer, few of the possible alternative goods are chosen. When faced with a set of similar products, customers are often observed to only purchase one product. For these situations, utility functions must be chosen which allow for corner solutions. This means that marginal utility must remain finite as consumption approaches zero (at least for some of the goods). In other words, the indifference curves cannot be tangent to the axes.

For situations in which consumers are only observed to choose one alternative, a linear utility function is a reasonable place to start.

$$\begin{aligned} &\max_x \; U(x) = \psi' x \\ &\text{s.t. } p' x = E \end{aligned} \tag{4.4.1}$$

For this functional form, marginal utility is constant and the optimal solution is to choose the alternative with the highest ratio of marginal utility to price (ψ_j / p_j) and consume $x = E/p_j$ units. In this situation, if we observe p and E, then consumer choice will be deterministic. The standard random utility approach is to assume that there are factors affecting the marginal utility which are not observed by the researcher but are known to the consumer. These factors could be time varying (i.e. consumer inventory or promotional activities) or represent taste differences between consumers. Since marginal utility must be positive, it is natural to consider a multiplicative specification for inclusion of these omitted unobservables.

$$\psi = \bar{\psi} e^{\varepsilon} \tag{4.4.2}$$

If we take logs, allow $\bar{\psi}$ to be a function of observable covariates such as consumer characteristics and product attributes, and subtract log price, we obtain a model similar to the standard multinomial models.

$$\ln \psi - \ln p = \bar{\psi} - \ln p + \varepsilon = \tilde{X}\beta - \ln p + \varepsilon \tag{4.4.3}$$

$\tilde{X}$ is a p x k matrix containing the revelevant covariates. Letting $z = \ln\psi - \ln p$ and $X = \left[\tilde{X} \quad \ln p \right]$, we have the latent variable formulation of a multinomial choice model in (4.2.1). Alternative j is chosen if $z_j = \max_i \left(z_i \right)$.

To complete the model, we must specify the distribution of the latent component of marginal utility. First, consider iid errors with a given scale, σ. In (4.4.3), σ is identified as the price coefficient is set to -1. Alternatively, we could introduce a price coefficient. If we allow a free price coefficient, we must fix the scale of the marginal utility errors to avoid the scale invariance identification problem. This means that we can interpret a price coefficient as $1/\sigma$. As the variance of the marginal utility errors grows, the influence of price on the probabilities of choice declines since the factors other than price are "larger" in scale. If we allow ε to have a multivariate normal distribution, we obtain the standard MNP model. The standard multinomial logit model can be derived by assuming that errors are iid extreme value type I. Demand theory provides some guidance as to the sort of covariates which might be include as well as the restriction that the price coefficient is negative and the same across choice alternatives.

4.4.1 A Non-homothetic Choice Model

The restrictive properties of the MN Logit model can be relaxed either by specifying a richer class of error distributions (as in the MNP model which avoids the IIA property) or by specifying a more flexible utility structure. The linear structure assumes constant marginal utility and no interactions (e.g. the marginal utility of alternative i is not affected by the consumption of other alternatives). One could argue that for choice data, constant marginal utility is appropriate in that we would need quantity information to estimate the curvature of the utility function. However, even with only access to choice information, the constant marginal utility assumption implies that price changes can only have substitution and not income effects (i.e. the utility function is homothetic). In the homothetic world, as greater amounts are allocated to expenditure on the products under study, we continue to purchase the same brand. In many marketing applications, there are products of differing quality levels. We observe consumers trading up to high quality models as expenditure increases or during sales. The opposite phenomena of "trading down" to lower quality brands as expenditure decreases or as lower quality brands are discounted does not occur as frequently. This phenomenon has been dubbed "asymmetric switching." The standard homothetic logit or probit models cannot exhibit asymmetric switching.

A non-homothetic choice model can be specified by retaining the linear structure of utility but assuming that marginal utility is a function of the overall level of attainable utility. This retains the assumption of no interactions in utility but allows for a non-trivial income effect which occurs either across consumers as consumers allocate a greater expenditure to the product category or as the category becomes "cheaper" due to price reductions on some of the items. One convenient specification which defines a valid utility function is given below.

$$\psi_i(u) = \exp\{\alpha_i - k_i u\} \tag{4.4.4}$$

In (4.4.4), the marginal utility of the ith product is a function of the maximum attainable level of utility. This defines the utility function implicitly. k_i governs the rate at which marginal utility changes as the level of attainable utility increases. If $k_1 < k_2$, then, as the budget increases, consumers will tend to purchase product 1 more than 2. The utility function for this case has a set of linear but rotating indifference curves. In this sense, we can regard product 1 as a superior product that provides a higher marginal utility for consumers willing to allocate a greater real expenditure to the product category.

Since the indifference curves are linear, consumers will choose only one of the products in the category. As in the standard linear utility model, consumers choose the alternative which provides the highest level of attainable utility conditional on the level of expenditure for the category, E, and vector of prices, p. That is, consumers find the maximum of $\{ u^1 , u^2 , \ldots, u^p \}$. u^i solves

$$u^i = \psi_i\left(u^i\right)E/p_i \tag{4.4.5}$$

Taking logs of both sides, we can write this as the implicit solution to an equation of the form $\ln(y) = C_1 - C_2\, y$. If C_2 is positive, this equation has an implicit solution which can be easily calculated by Newton's method. To complete this model, we must specify a random utility error as well as an expenditure function. In Allenby and Rossi (1991), an expenditure function was specified as

$$\ln E = \gamma' z \tag{4.4.6}$$

z is a vector covariates. In Allenby and Rossi, z includes a price index for the product category. This is a somewhat ad hoc solution which avoids the specification of a bivariate utility function over all other goods and this product category. In addition, one might be tempted to make the α and k parameters a function of demographic variables. However, since marginal utility is specified as an implicit non-linear function of α and k, we would not be able to write the non-homothetic model in form of (4.4.3).

If we add a standard extreme value type I error to the implicitly defined marginal utility, we obtain a logit model in which the probability of choice is given by

$$\Pr\left(i\right) = \frac{\exp\left(\tau v_i\right)}{\sum_j \exp\left(\tau v_j\right)} \tag{4.4.7}$$

where $v_j = \alpha_j - k_j u^j - \ln p_j$. τ is the scale parameter of the extreme value distribution. Note that u is also a function of E, p_j , α_j and k_j. Thus, we can write $v_j = f\left(\alpha_j, k_j \middle| E, p_j\right)$. `llnhlogit` in *`bayesm`* evaluates the log-likelihood for this model.

4.4.2 Demand for Discrete Quantities

The application of discrete choice models to packaged goods requires researchers to adjust their models to accommodate demand quantities. Packaged goods are often available in multiple sizes, and it is not appropriate to treat the various package sizes as independent alternatives since the same good is contained in each of the packages. While it is possible to estimate discrete choice models that allow for dependence among the alternatives, additional restrictions on the coefficients are needed so that parameter estimates from the model conform to economic theory. For example, a discrete choice model calibrated on soft drink purchases would need to impose ordinal restrictions on the intercepts so that the utility of 6-pack, 12-pack and 18-pack offerings would reflect diminishing marginal returns to quantity.

Quantity can be incorporated into models of consumer demand by imbedding a utility function for a discrete choice model into a utility function that relates the product class to an outside good. For example, consider the Cobb-Douglas utility function:

$$\ln U(x,z) = \alpha_0 + \alpha_x \ln U(x) + \alpha_z \ln(z) \qquad (4.4.8)$$

where x = (x_1, …, x_K) is the vector of the amount of each alternative (i.e., brand) purchased, K represents the number of brands in the product class, z represents the amount of the outside good purchased, and U(x) denotes a sub-utility function. The sub-utility function, U(x), can be specified as the linear function in equation (4.4.1) or the non-homothetic function in equation (4.4.4).

Maximizing (4.4.8) subject to a budget constraint leads to a vector of demand (x,z) that, in general, is a mixture of corner and interior solutions. However, as discussed in case study 1, "A Choice Model for Packaged Goods," the utility maximizing solution will always be a corner solution in x when per-unit price schedules are concave, e.g. the cost per ounce of a 6-pack of is greater than the cost per ounce of a 12-pack. When this occurs, the choice probability of observing quantity x_i is:

$$\Pr(x_i) = \frac{\exp\left[\ln(\psi_i) + \ln(x_i) + (\alpha_z / \alpha_x)\ln(E - p_i(x_i))\right]}{\sum_{k=1}^{K} \exp\left[\ln(\psi_k) + \ln(x_k) + (\alpha_z / \alpha_x)\ln(E - p_k(x_i))\right]} \tag{4.4.9}$$

where ψ_i is the marginal utility of brand i, E is the budgetary allotment, $p_i(x_i)$ is the price of x_i units of brand i, and x_k is the quantity of brand k that maximizes equation (4.4.8).

4.4.3 Demand for Variety

In some product categories, consumers are observed to purchase a subset of products in the category. For example, consumers purchase multiple varieties of soft drinks or yogurts. The standard multinomial models have zero likelihood for this sort of consumer behavior as the choice options are regarded as mutually exclusive. On the other hand, many common utility specifications are designed to give rise to strictly interior solutions in which all products in the category are purchased. What is needed is a demand system which can give rise to a mixture of corner and interior solutions. This can be achieved by translating a utility function so that its indifference curves intersect the axes with finite slope. One simple additive structure is given by

$$\bar{U}(x) = \sum_j \psi_j \left(x_j + \gamma_j\right)^{\alpha_j} \tag{4.4.10}$$

The $\{\gamma_j\}$ parameters serve to translate an additive power utility to admit the possibility of corner solutions. The utility function also exhibits curvature or diminishing marginal utility which allows for the possibility of "wear-out" in the consumption of a particular variety. The utility in (4.4.10) is an additive, but non-linear utility function. (4.4.10) defines a valid utility function under the restrictions that $\psi_j > 0$ and $0 < \alpha_j \leq 1$.

This utility specification can accommodate a wide variety of situations, including the purchase of a large number of different varieties as well as purchases where only one variety is selected. If a particular variety has a high value of ψ_j and a value of α_j close to one, then we would expected to see purchases of large quantities of only one variety (high baseline preference and low satiation). On the other hand, small values of α imply a high satiation rate, we expect to see multiple varieties purchased if the ψ's are not too different.

To develop a statistical specification, we follow a standard random utility approach and introduce a multiplicative normal error into marginal utility:

$$\text{(4.4.11)} \qquad \ln\left(U_j\right) = \ln\left(\bar{U}_j\right) + \varepsilon_j \quad \varepsilon \sim N\left(0, \Sigma\right)$$

where $\bar{U}_j$ is the derivative of the utility function in (4.4.10) with respect to x_j. We use a log-normal error term to enforce positivity of marginal utility. We specify a full covariance matrix for the random marginal utility errors. In some applications, it may be difficult to identify this covariance matrix. Further restrictions may be necessary. Even the assumption that Σ is the identity matrix is not necessarily too restrictive as we have specified a log-normal distribution of marginal utility errors which exhibits heteroskedasticity of a reasonable form. However, this is largely an empirical matter.

We derive the demand system for the set of goods under study conditional on the expenditure allocation to this set of goods. In the random utility approach, it is assumed that the consumer knows the value of ε and that this represents omitted factors which influence marginal utility but are not observable to the data analyst. If we derive the optimal demand by maximizing utility subject to the budget constraint and conditional on the random utility error, we define a mapping from p, E and ε to demand. Assuming a distribution for ε provides a basis for deriving the distribution of optimal demand, denoted x*. There are two technical issues in deriving the distribution of demand: 1). Optimal demand is a non-linear function of ε and requires use of change-of-variable calculus and 2). The possibility of corner solutions means that there are point masses in the distribution of demand and, thus, the distribution of demand will be a mixed discrete-continuous distribution. Computing the size of these point masses involves integrating the normal distribution of ε over rectangular regions of R^m.

To solve for optimal demand, we form the Lagrangian for the problem and derive the standard Kuhn-Tucker first order conditions. It is important to remember in the utility function specified in (4.4.11), $\bar{U}$ is only the deterministic part of utility (that observed by us) and that the consumer maximizes U which includes the realization of the random utility errors. The Lagrangian is given by

$$U(x)-\lambda(p'x-E)$$

Differentiating the Lagrangian gives the standard Kuhn-Tucker first order conditions:

$$\bar{U}_j e^{\varepsilon_j}-\lambda p_j=0 \qquad x_j^*>0$$

$$\bar{U}_j e^{\varepsilon_j}-\lambda p_j<0 \qquad x_j^*=0$$

x* is the vector of optimal demands for each of the m goods under consideration. Dividing by price and taking logs, the Kuhn-Tucker conditions can be re-written as:

$$\begin{array}{ll} V_j\left(x_j^* \mid p\right)+\varepsilon_j=\ln\lambda & \text{if } x_j^*>0 \\ V_j\left(x_j^* \mid p\right)+\varepsilon_j<\ln\lambda & \text{if } x_j^*=0 \end{array} \tag{4.4.12}$$

where λ is Lagrange multiplier and $V_j\left(x_j^* \mid p\right) = \ln\left(\psi_j \alpha_j \left(x_j^* + \gamma_j\right)^{\alpha_j - 1}\right) - \ln\left(p_j\right)$ j=1,….,m.

Optimal demand satisfies the Kuhn-Tucker conditions in (4.4.12) as well as the "adding-up" constraint that total $p'x^* = E$. 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. Without loss of generality, we assume that the first good is always purchased (one of the m goods must be purchased since we assume that E >0) and subtract condition, (4.4.12), for good 1 from the others. This reduces the dimensionality of the system of equations by one. (4.4.12) is now equivalent to:

$$\begin{aligned} \upsilon_j &= h_j\left(x^*, p\right) \quad \text{if } x_j^* > 0 \\ \upsilon_j &< h_j\left(x^*, p\right) \quad \text{if } x_j^* = 0 \end{aligned} \tag{4.4.13}$$

where $\nu_j = \varepsilon_j - \varepsilon_1$ and $h_j(x^*, p) = V_1 - V_j$ and j=2, …, m.

The likelihood for $x^* = \left(x_1^*, \ldots, x_m^*\right)'$ can be constructed by utilizing the p.d.f. of $\upsilon = (\upsilon_2, \ldots, \upsilon_m)'$, the Kuhn-Tucker conditions in (4.4.13), and the adding up constraint $p'x^* = E$. $\upsilon = (\upsilon_2, \ldots, \upsilon_m)' \sim N\left(0, \Omega\right)$. $\Omega = A\Sigma A'$ where $A = \left[-\iota \quad I_{m-1}\right]$. Given that corner solutions will occur with non-zero probability, the distribution of optimal demand will have a mixed discrete-continuous distribution with lumps of probability corresponding to regions of ε which imply corner solutions. Thus, the likelihood function will have a density component corresponding to the goods with non-zero quantities and a mass function corresponding to the corners in which some of the goods will have zero optimal demand. The probability that n of the m goods are selected is equal to:

(4.4.14)
$$\begin{aligned}&P\left(x_i^* > 0 \text{ and } x_j^* = 0;\ i = 2,\ldots,n \text{ and } j = n+1,\ldots m\right)\\ &\quad = \int_{-\infty}^{h_m}\cdots\int_{-\infty}^{h_{n+1}} \phi(h\upsilon_n, \upsilon_{n+1},\ldots,\upsilon_m \mid 0,\ \Omega)\,|\,J\,|\,d\upsilon_{n+1}\cdots d\upsilon_m\end{aligned}$$

where $\phi(\cdot)$ is normal density, $h_j = h_j(x^*, p)$, and J is the Jacobian,

$$J_{ij} = \frac{\partial h_{i+1}(x^*;p)}{\partial x_{j+1}^*} \qquad i, j = 1,\ldots., n-1\,.$$

We should note that the adding up constraint, p'x = E, makes this Jacobian non-diagonal as we can always express the demand for the "first" good with non-zero demand as a function of the other demands.

The intuition behind the likelihood function in (4.4.14) can be obtained from the Kuhn-Tucker conditions in (4.4.13). For goods with non-zero demand, the first condition in (4.4.13) means that optimal demand is an implicitly defined non-linear function of ε given by h(). We use the change-of-variable theorem to derive the density of x* (this generates the Jacobian term). For goods not purchased, the second Kuhn-Tucker condition defines a region of possible values of ν which are consistent with this specific corner solution. The probability that these goods have zero demand is calculated by integrating the normal distribution of ν over the appropriate region.

If there are only corner solutions with one good chosen, our model collapses to a standard choice model. The probability that only good one is chosen is given by

$$\begin{aligned}&P\left(x_j^* = 0,\ j = 2,\ldots,m\right)\\ &\quad = \int_{-\infty}^{h_m}\cdots\int_{-\infty}^{h_2} \phi\left(\nu_2,\ldots,\nu_m\right)\,d\nu_2\cdots d\nu_m\end{aligned}$$

Similarly, we can derive the distribution of demand for the case in which all goods are at an interior solution.

$$P\left(x_i^* > 0;\ i = 2,\ldots,m\right)$$
$$= \varphi\left(h_2,\ldots,h_m \,|\, 0,\Omega\right)|J|$$

The joint distribution of $(x_1^*,....,x_m^*)'$ in (4.4.14) can be evaluated by noting that it can be factored into discrete and continuous parts. In evaluating the likelihood we transform (7) to the product of two factors as follows. By partitioning $\upsilon = (\upsilon_2,.....,\upsilon_m)'$ into $\nu_a = (\upsilon_2,..,\upsilon_n)'$ and $\nu_b = (\upsilon_{n+1},..,\upsilon_m)'$ such that

$$\begin{bmatrix} \nu_a \\ \nu_b \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix},\ \begin{bmatrix} \Omega_{aa} & \Omega_{ab} \\ \Omega_{ba} & \Omega_{bb} \end{bmatrix}\right)$$

ν_a and $\nu_b \,|\, \nu_a$ are normally distributed, then $\nu_a \sim \text{MVN}(0,\ \Omega_{aa})$ and $\nu_b \,|\, \nu_a = h_a \sim \text{MVN}(\mu,\Sigma)$ where $\mu = \Omega_{ba}\Omega_{aa}^{-1}h_a$, $\Sigma = \Omega_{bb} - \Omega_{ba}\Omega_{aa}^{-1}\Omega_{ab}$, and $h_a = (h_2,....,h_n)'$. Then, (4.4.14) can be rewritten as the product of two factors:

$$\begin{aligned} &P\left(x_i^* > 0 \text{ and } x_j^* = 0;\ i = 2,\ldots,n \text{ and } j = n+1,\ldots,m\right) = \\ &\phi_{\nu_a}\left(h_2,\ldots,h_n \,|\, 0,\Omega_{aa}\right)|J| \int_{-\infty}^{h_m}\cdots\int_{-\infty}^{h_{n+1}} \phi_{\nu_b|\nu_a}\left(\nu_{n+1},\ldots,\nu_m \,|\, \mu,\Sigma\right)d\nu_{n+1}\cdots d\nu_m \end{aligned} \quad (4.4.15)$$

We use the GHK simulator (Chapter 2, section 10 or Keane, 1994, Hajivassilious, McFadden and Ruud, 1996) to evaluate the multivariate normal integral in (4.4.15) .

In the case study, "Modeling Consumer Demand for Variety," we will apply a heterogeneous version of this model to data on purchase of yogurt varieties. We will also discuss the R implementation of this model.

The additive utility model used does not include any interactions in the utility function. That is, the marginal utility of consumption of good i does not depend on the consumption level of other goods. In particular, additive utility specifications impose the restriction that all goods are substitutes, ruling out complementarity. Gentzkow (2005) includes utility interaction in a choice model which allows for the possibility that two goods are complements.[6] He applies this formulation to purchase data on the print and on-line versions of newspapers. His results suggest that there are complementarities between the print and on-line versions of newspapers.

[6] We note that with only two goods there can only be substitution in demand. However, Gentzkow includes the usual outside alternative.

Figure 4.1

MNP Sampler for Simulated Data with p=6

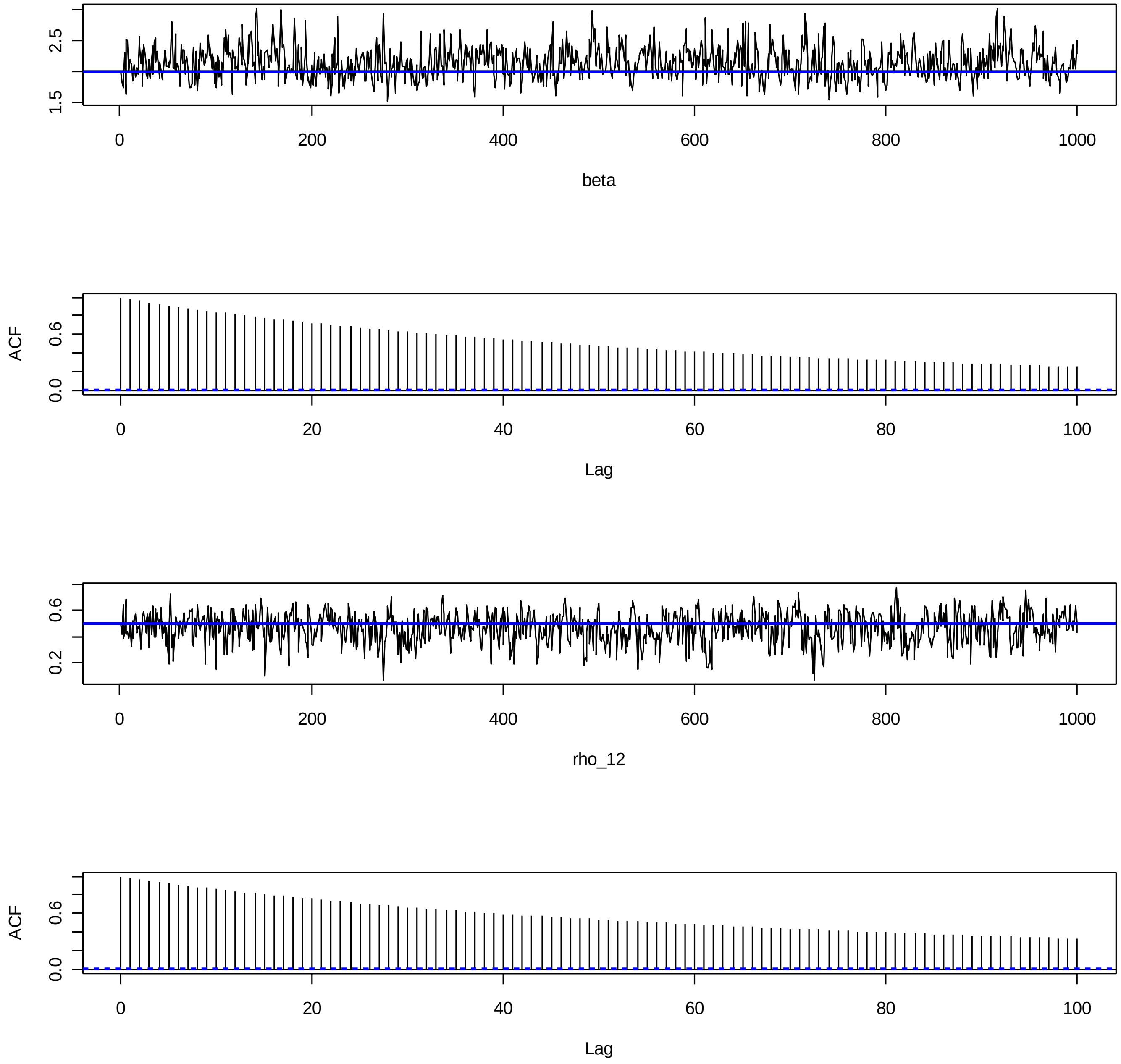


Figure 4.2

Understanding the Correlation Between Latent Variable and Binary Logit Parameters

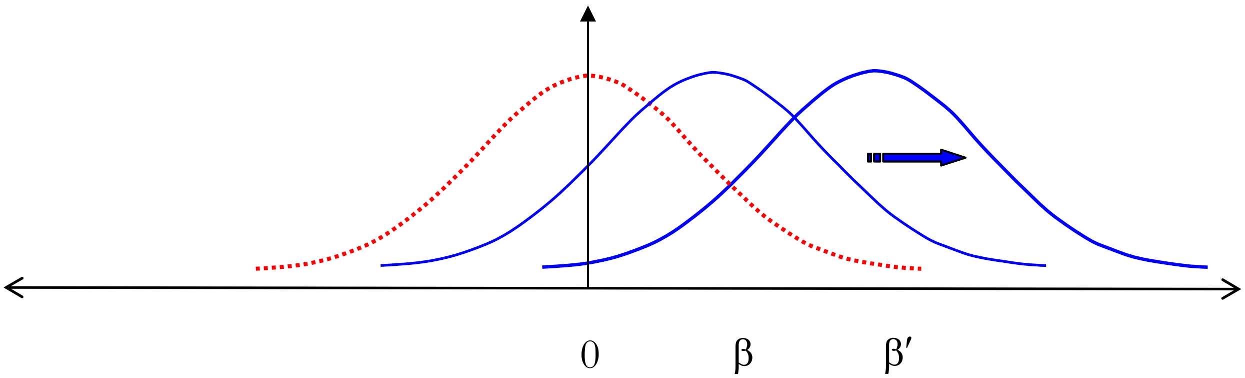


Figure 4.3

Correlation between β and latents

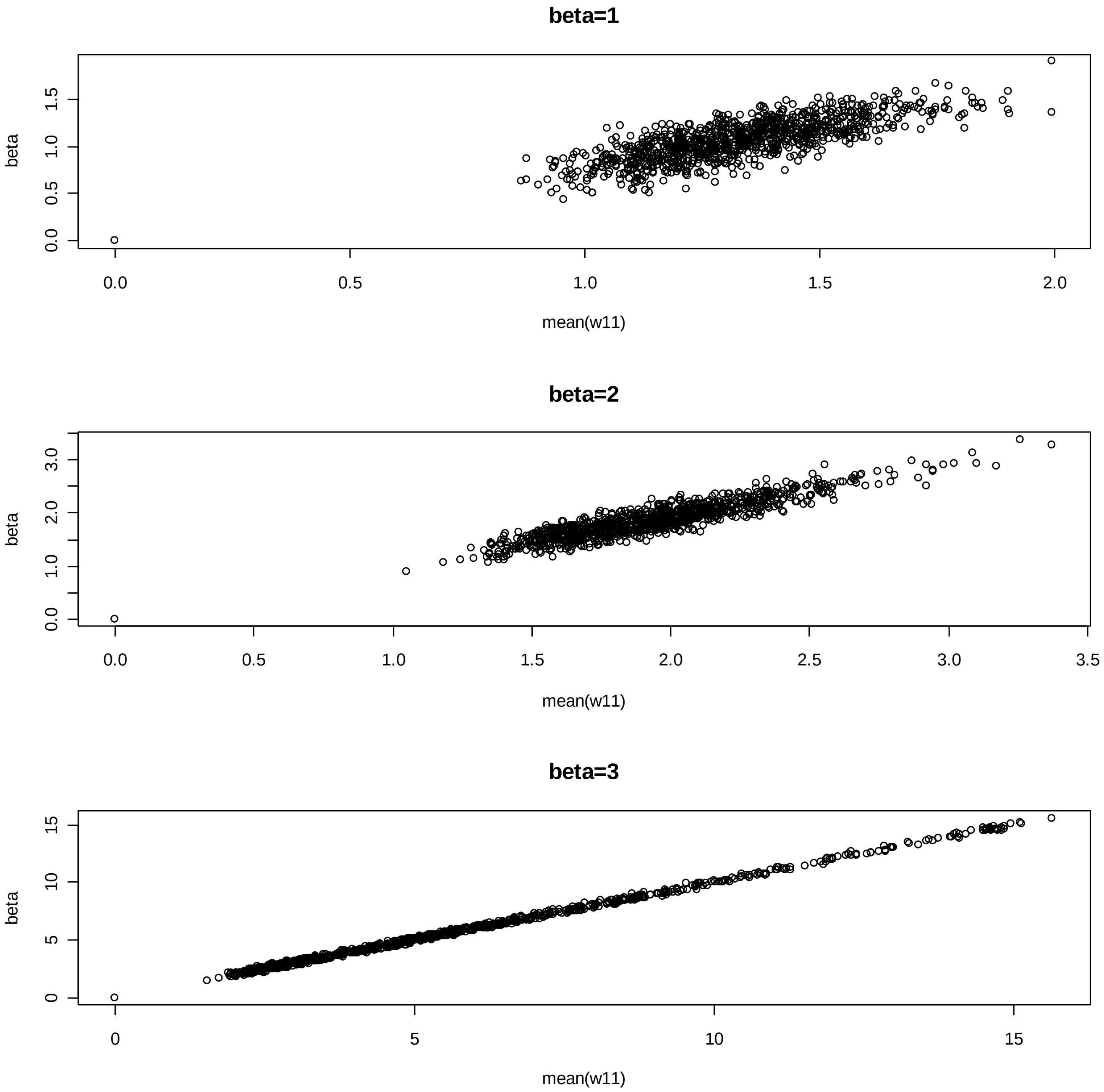
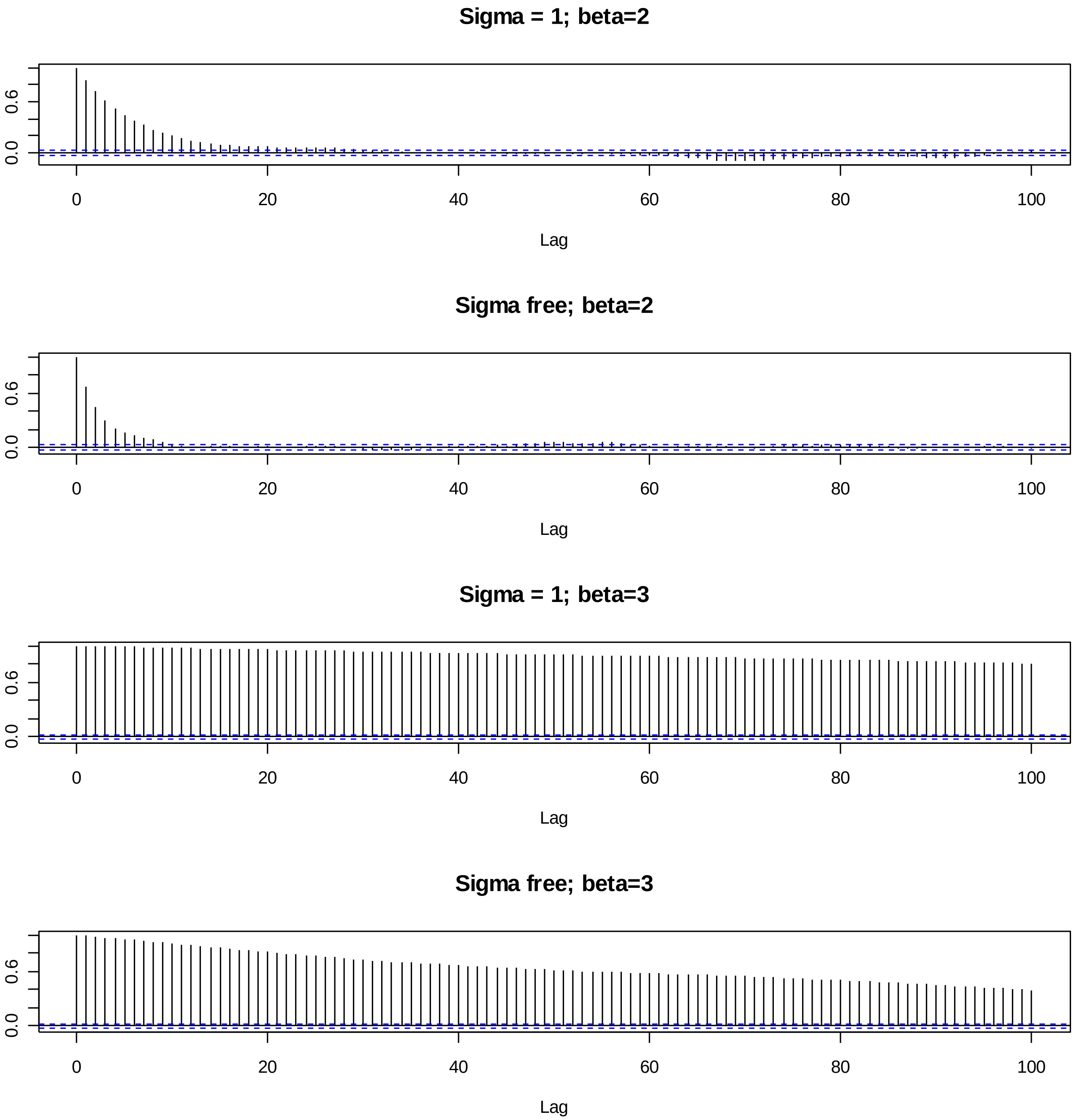


Figure 4.4

Comparison of NID and ID MNP Samplers



Sigma = 1; beta=2

Sigma free; beta=2

Sigma = 1; beta=3

Sigma free; beta=3

Figure 4.5

Regions of Integration for MNP Model

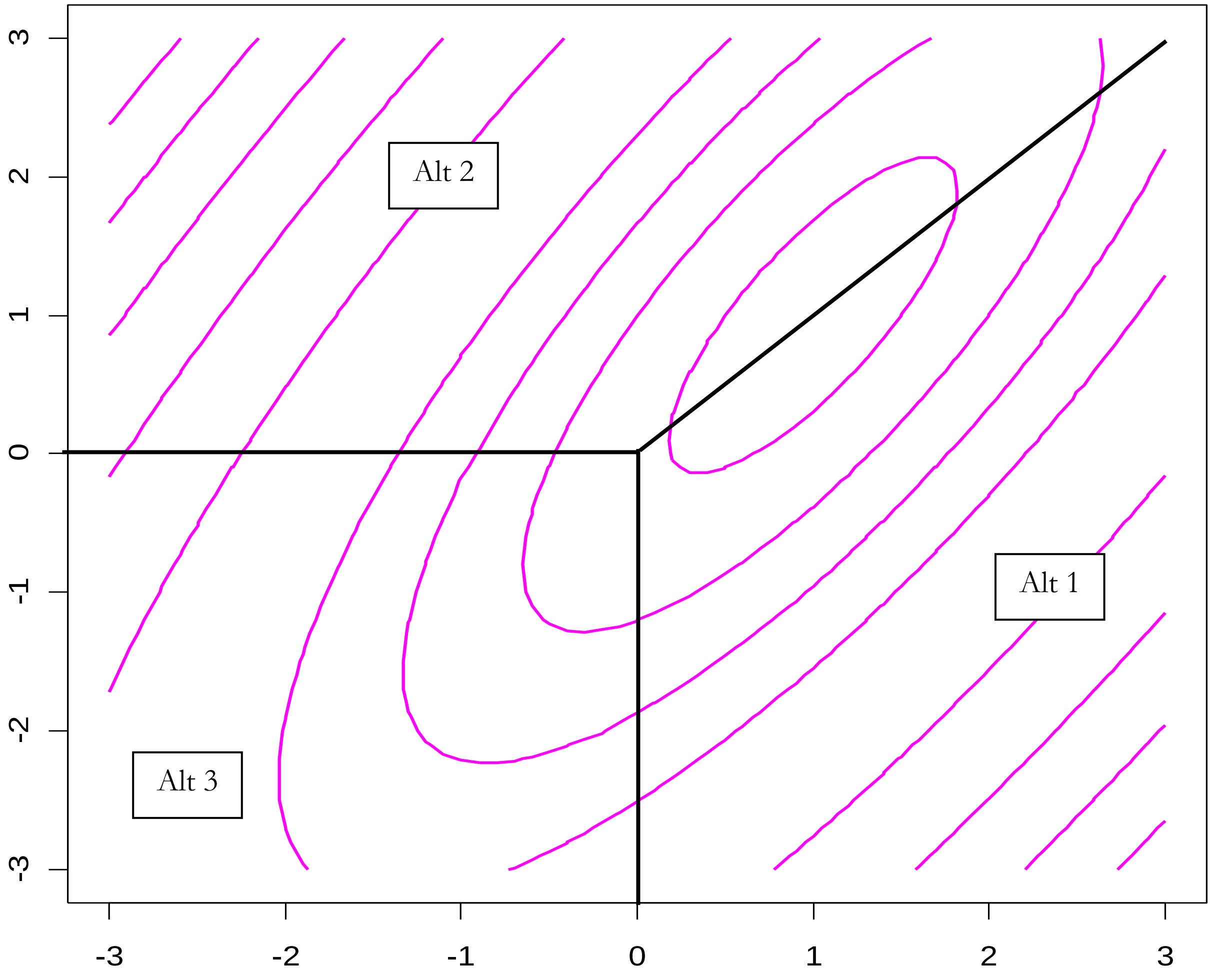


Figure 4.6
Multivariate Probit Model Intercepts:
Scotch Survey Data

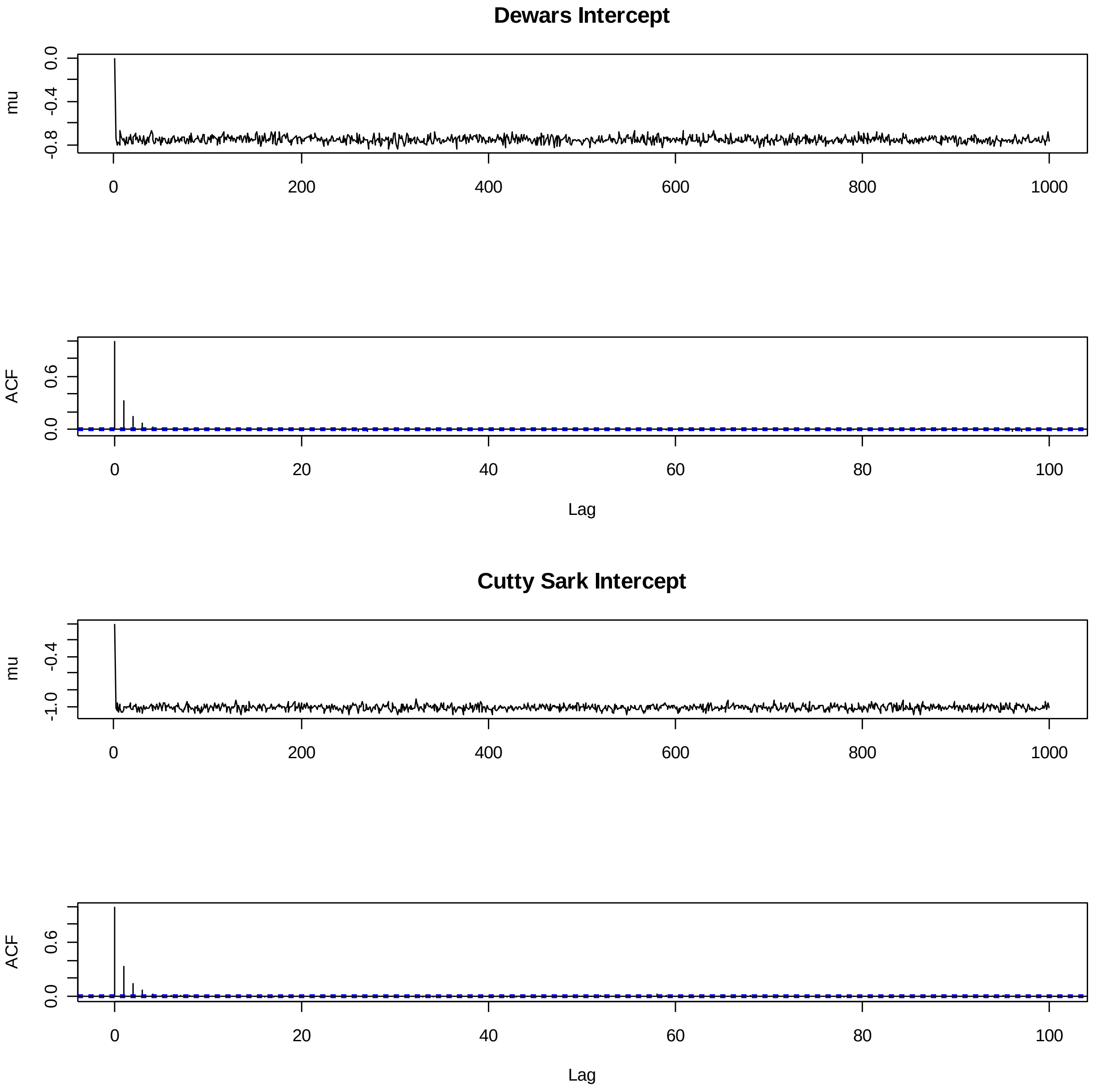
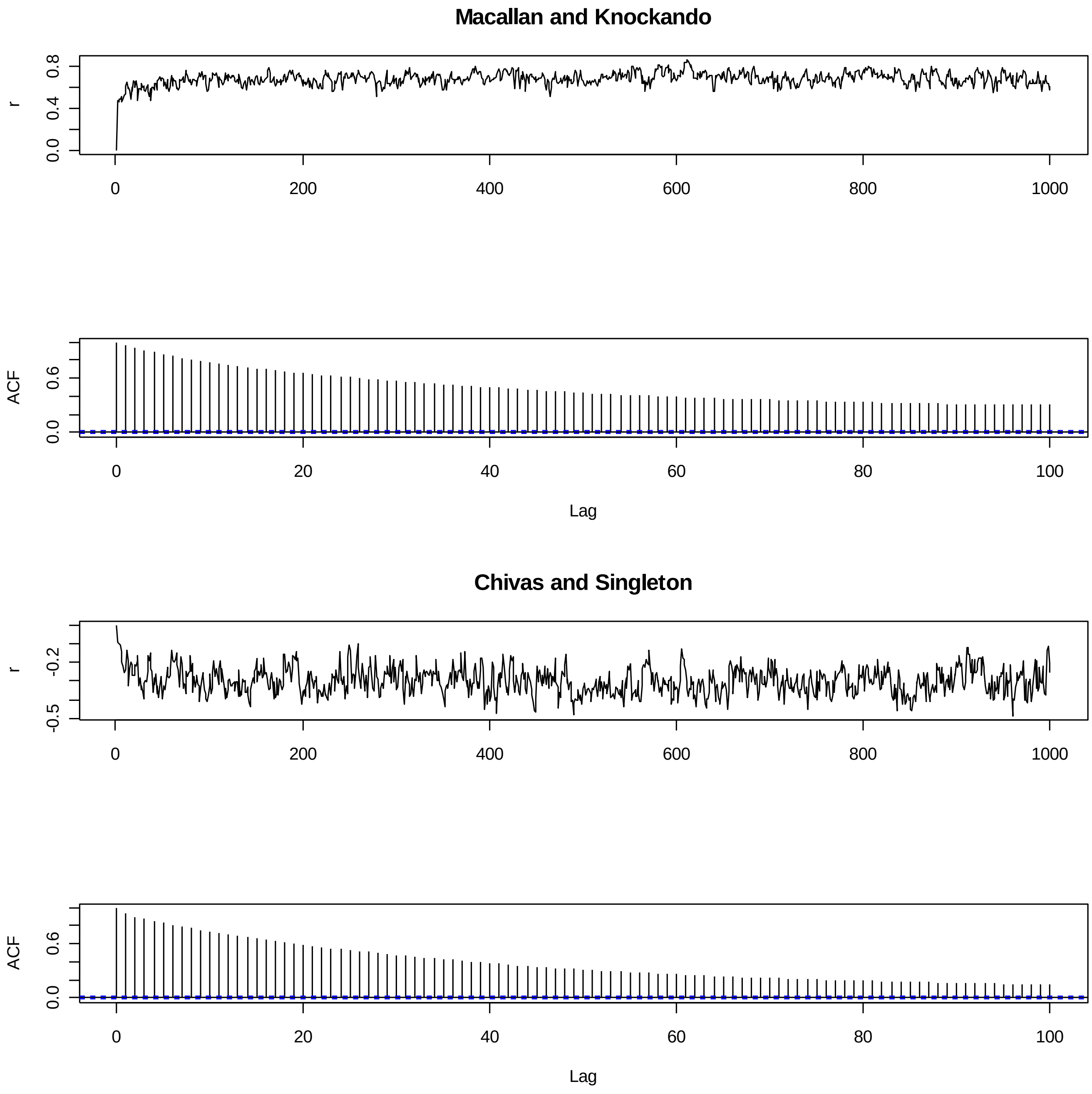
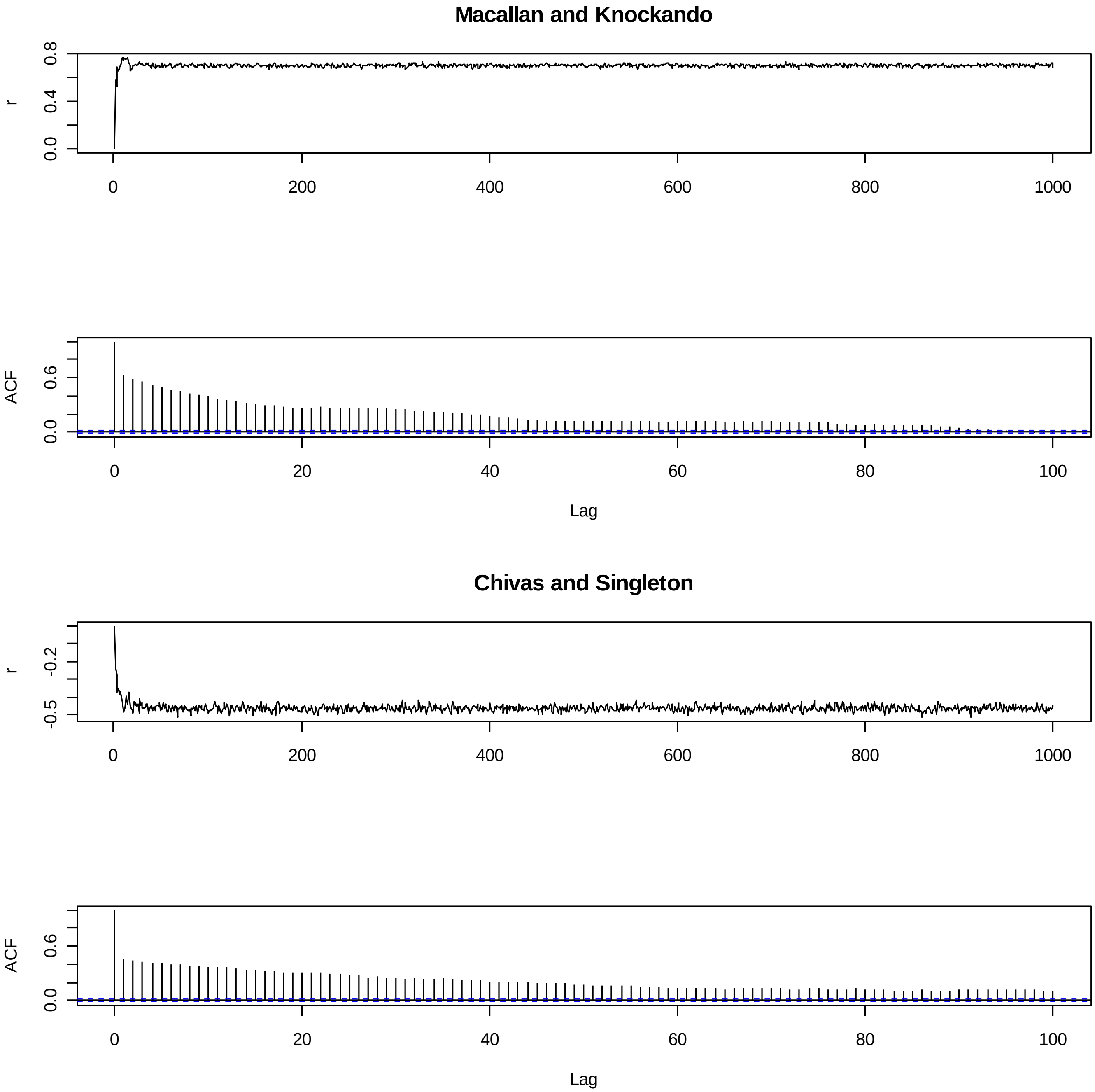


Figure 4.7
Multivariate Probit Model Correlations:
Scotch Survey Data


Macallan and Knockando

Chivas and Singleton

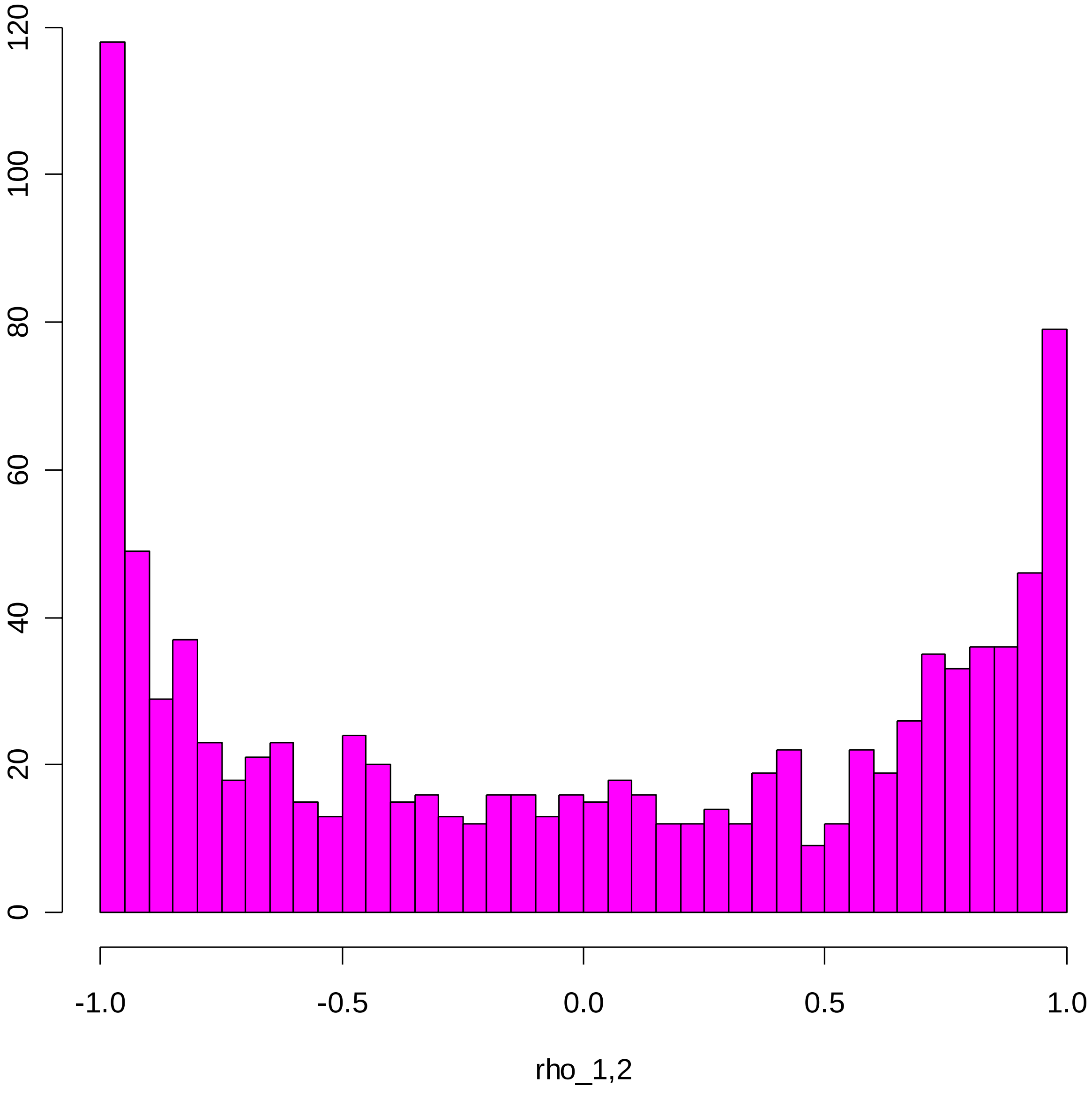
Figure 4.8
Multivariate Probit Model Correlations:
Scotch Survey Data with Improper Prior on Σ



Macallan and Knockando

Chivas and Singleton

Figure 4.9
Prior Distribution of a Correlation Coefficient


rho_1,2