Chapter 5 Hierarchical Models for Heterogeneous Units

Using this Chapter

This chapter provides a comprehensive treatment of hierarchical models. Hierarchical models are designed to measure differences between units using a particular prior structure. Choice of the form of the hierarchical model (i.e. the form of the prior) as well as the MCMC algorithm to conduct inference are important questions. We explore a new class of hybrid MCMC algorithms that are customized or tuned to the posteriors for individual units. We also implement a mixture of normals prior for the distribution of model coefficients across units. We illustrate these methods in the context of a panel of household purchase data and a base or unit-level multinomial logit model. Those interested in the main points without technical details are urged to concentrate on sections 5.1, 5.2, 5.4, and 5.5.3.

One of the greatest challenges in marketing is to understand the diversity of preferences and sensitivities that exists in the market. Heterogeneity in preferences gives rise to differentiated product offerings, market segments and market niches. Differing sensitivities are the basis for targeted communication programs and promotions. As consumer preferences and sensitivities become more diverse, it becomes less and less efficient to consider the market in the aggregate. Marketing practices which are designed to respond to consumer differences require an inference method and model capable of producing individual or unit level parameter estimates. Moreover, optimal decision-making requires not-only point estimates of unit level parameters but also a characterization of the uncertainty in these estimates. In this chapter, we will show how Bayesian hierarchical approaches are ideal for these problems as it is possible to produce posterior distributions for a large numbers of unit-level parameters.

In contrast to this emphasis on individual differences, economists are often more interested in aggregate effects and regard heterogeneity as a statistical nuisance parameter problem which must be addressed but not emphasized. Econometricians frequently employ methods which do not allow for the estimation of individual-level parameters. For example, random coefficient models are often implemented through an unconditional likelihood

approach in which only hyper-parameters are estimated. Furthermore, the models of heterogeneity considered in the econometrics literature often restrict heterogeneity to subsets of parameters such as model intercepts. In the marketing context, there is no reason to believe that differences should be confined to the intercepts and, as indicated above, differences in slope coefficients are critically important. Finally, economic policy evaluation is frequently based on estimated hyper-parameters which are measured with much greater certainty than individual-level parameters. This is in contrast to marketing policies which often attempt to respond to individual differences that are measured less precisely.

This new literature emphasizing unit-level parameters is made possible by the explosion in the availability of disaggregate data. Scanner data at the household and store level is now commonplace. In the pharmaceutical industry, physician-level prescription data is also available. This raises both modeling challenges as well as major opportunities for improved profitability through decentralized marketing decisions that exploit heterogeneity. This new data comes in panel structure in which N, the number of units is large relative to T, the length of the panel. Thus, we may have a large amount of data obtained by observing a large number of decision units. For a variety of reasons, it is unlikely that we will ever have a very large amount of information about any one decision unit. Data describing consumer preferences and sensitivities to variables such as price are typically obtained through surveys or household purchase histories which yield very limited individual-level information. For example, household purchases in most product categories often total less than 12 per year. Similarly, survey respondents become fatigued and irritable when questioned for more than 20 or 30 minutes. As a result, the amount of data available for drawing inferences about any specific consumer is very small, although there may exist many consumers in a particular study.

The classical fixed-effects approach to heterogeneity has some appeal since it delivers the individual unit-level parameter estimates and does not require the specification of any particular probability distribution of heterogeneity. However, the sparseness of individual-level data renders this approach impractical. In many situations, incomplete household level data causes a lack of identification at the unit level. In other cases, the parameters are identified in the unit-level likelihood but the fixed effects estimates are measured with huge uncertainty which is difficult to quantify using standard asymptotic methods.

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 chapter, we will introduce hierarchial models for general unit level models and apply these ideas to a hierarchical logit setting. Recognizing the need for flexibility in the prior, we will expand the set of priors to include mixtures of normal distributions.

5.1 Heterogeneity and Priors

A useful general structure for disaggregate data is a panel structure in which the units are regarded as independent conditional on unit level parameters (see Case Study 2, "Modeling Interdependent Consumer Preferences," for an example which relaxes this assumption). Given a joint prior on the collection of unit level parameters, the posterior distribution can be written as follows:

$$p(\theta_1,...,\theta_N|y_1,...,y_m) \propto \left[\prod_i p(y_i|\theta_i)\right] \times p(\theta_1,...,\theta_m|\tau) \tag{5.1.1}$$

The term in brackets is the conditional likelihood and the rightmost term is the joint prior with hyperparameter, τ. In many instances, the amount of information available for many

of the units is small. This means that the specification of the functional form and hyper-parameter for the prior may be important in determining the inferences made for any one unit. A good example of this can be found in choice data sets in which consumers are observed to be choosing from a set of products. Many consumers ("units") do not choose all of the alternatives available during the course of observation. In this situation, most standard choice models don't have a bounded maximum likelihood estimate (the likelihood will asymptote in a certain direction in the parameter space). For these consumers, the prior is, in large part, determining the inferences made.

Assessment of the joint prior for $(\theta_1,...,\theta_m)$ is difficult due to the high dimension of the parameter space and, therefore, some sort of simplification of the form of the prior is required. One frequently employed simplification is to assume that, conditional on the hyperparameter, $(\theta_1,...,\theta_m)$ are a priori independent.

(5.1.2) $$p(\theta_1,...,\theta_m|y_1,...,y_m) \propto \prod_i p(y_i|\theta_i)p(\theta_i|\tau)$$

This means that inference for each unit can be conducted independently of all other units *conditional* on τ. This is the Bayesian analogue of fixed effects approaches in classical statistics.

The specification of the conditionally independent prior can be very important due to the scarcity of data for many of the units. Both the form of the prior and the values of the hyperparameters are important and can have pronounced effects on the unit-level inferences. For example, it is common to specify a normal prior, $\theta_i \sim N(\bar{\theta}, V_\theta)$. The normal form of this prior means that influence of the likelihood for each unit may be attenuated for likelihoods centered far away from the prior. That is, the thin tails of the normal distribution diminish the influence of outlying observations. In this sense, the

specification of a normal form for the prior, whatever the values of the hyper-parameters, is far from innocuous.

Assessment of the prior hyperparameters can also be challenging in any applied situation. For the case of the normal prior, some relatively diffuse prior may be a reasonable default choice. Allenby and Rossi (1993) use a prior based on a scaled version of the pooled model information matrix. The prior covariance is scaled back to represent the expected information in one observation to insure a relatively diffuse prior. Use of this sort of normal prior will induce a phenomenon of "shrinkage" in which the Bayes estimates (posterior means) $\left\{\tilde{\theta}_i = E\left[\theta_i \middle| data_i, prior\right]\right\}$ will be clustered more closely to the prior mean than the unit-level maximum likelihood estimates $\left\{\hat{\theta}_i\right\}$. For diffuse prior settings, the normal form of the prior will be responsible for the shrinkage effects. In particular, outliers will be "shrunk" dramatically toward the prior mean. For many applications, this is a very desirable feature of the normal form prior. We will "shrink" the outliers in towards the rest of the parameter estimates and leave the rest pretty much alone.

5.2 Hierarchical Models

In general, however, it may be desirable to have the amount of shrinkage induced by the priors driven by information in the data. That is, we should "adapt" the level of shrinkage to the information in the data regarding the dispersion in $\{\theta_i\}$. If, for example, we observe that the $\{\theta_i\}$ are tightly distributed about some location or that there is very little information in each unit level likelihood, then we might want to increase the tightness of the prior so that the shrinkage effects are larger. This feature of "adaptive shrinkage" was the original motivation for work by Efron and Morris (1975) and others on empirical Bayes approaches in which prior parameters were estimated. These empirical Bayes approaches are

an approximation to a full Bayes approach in which we specify a second stage prior on the hyper-parameters of the conditional independent prior. This specification is called a Hierarchical Bayes Model and consists of the unit level likelihood and two stages of priors.

Likelihood: $p(y_i|\theta_i)$

First-stage prior: $p(\theta_i|\tau)$

Second-stage prior: $p(\tau|h)$

The joint posterior for the hierarchical model is given by

$$p(\theta_1,\ldots,\theta_m,\tau|y_1,\ldots,y_m,h) \propto \left[\prod_i p(y_i|\theta_i)p(\theta_i|\tau)\right] \times p(\tau|h)$$

In the hierarchical model, the prior induced on the unit level parameters is not an independent prior. The unit level parameters are conditionally, but not unconditionally, a priori independent.

$$p(\theta_1,\ldots,\theta_m|h) = \int \prod_i p(\theta_i|\tau)p(\tau|h)d\tau$$

If, for example, the second-stage prior on τ is very diffuse, the marginal priors on the unit-level parameters, θ_i, will be highly dependent as each parameter has a large common component. Improper priors on the hyper-parameters are extremely dangerous not only because of their extreme implications for some marginals of interest as we have seen in chapter 4 but also because the posterior may not be proper. As Hobert and Casella (1996) point out, it is possible to define an MCMC method for a hierarchical model which does not have any posterior as its invariant distribution for the case of improper priors.

The first-stage prior (or random effect distribution) is often taken to be a normal prior. Obviously, the normal distribution is a flexible distribution with easily interpretable parameters. In addition, we can increase the flexibility of this distribution using a mixture

of normals approach as outlined in section 5.5. We can easily incorporate observable features of each unit by using a Multivariate Regression specification.

$$\theta_i = \Delta' z_i + u_i \quad u_i \sim N(0, V_\theta)$$
(5.2.1) or
$$\Theta = Z\Delta + U$$

Θ is an m x k matrix whose rows contain each of the unit level parameter vectors. Z is an m x n_z matrix of observations on the n_z covariates which describe differences between units with row z_i. $\theta_i \sim N(\bar{\theta}, V_\theta)$ is a special case of (5.2.1) where Z is a vector of ones with length equal to the number of units. Given the Θ array, draws of Δ and V_θ can be accomplished using either a Gibbs sampler or direct draws for the Multivariate Regression model as outlined in section 2.12 and implemented in the ***bayesm*** function, `rmultireg`.

The hierarchical model specifies that both prior and sample information will be used to make inferences about the common parameter, τ. For example, in normal prior, $\theta_i \sim N(\bar{\theta}, V_\theta)$, the common parameters provide the location and the spread of the distribution of θ_i. Thus, the posterior for the θ_i will reflect a level of shrinkage inferred from the data. It is important to remember, however, that the normal functional form will induce a great deal of shrinkage for outlying units even if the posterior of V_θ is centered on large values.

In classical approaches to these models, the first-stage prior is called a random effects model and is considered part of the likelihood. The random effects model is used to average the conditional likelihood to produce an unconditional likelihood which is a function of the common parameters alone.

$$\ell(\tau) = \prod_i \int p(y_i|\theta_i) p(\theta_i|\tau) d\theta_i$$

In the classic econometric literature, much is made of the distinction between random coefficient models and fixed effect models. Fixed effect models are considered "non-parametric" in the sense that there is no specified distribution for the θ_i parameters[1]. Random coefficient models are often considered more efficient but subject to specification error in the assumed random effects distribution, $p(\theta_i|\tau)$. In a Bayesian treatment, we see that the distinction between these two approaches is in the formulation of the joint prior on $\{\theta_1,\ldots,\theta_m\}$. A Bayesian "fixed effects" approach specifies independent priors over each of the unit level parameters while the "random effects" approach specifies a highly dependent joint prior.

The use of a hierarchical model for prediction also highlights the distinction between various priors. A hierarchical model assumes that each unit is a draw from a "super-population" or that the units are exchangeable (conditional, perhaps, on some vector of co-variates). This means that if we want to make a prediction regarding a new unit we can regard this new unit as drawn from the same population. Without the hierarchical structure, all we know is that this new unit is different and have little guidance as to how to proceed.

5.3 Inference for Hierarchical Models

Hierarchical models for panel data structures are ideally suited for MCMC methods. In particular, a "Gibbs" style Markov chain can often be constructed by considering the basic two sets of conditionals:

1. $\theta_i|\tau, y_i$

[1] Classical inference for fixed effects models faces a fundamental conundrum: more time series observations are required for application of asymptotic theory (which is needed for non-linear models). However, we invariably have a short panel. Various experiments in which both the number of cross-sectional units and the time dimension increase are unpersuasive. While we might accept asymptotics that allow only N to increase to infinity, we are unlikely ever to see T increase as well. But most importantly, we avoid this altogether in the Bayesian approach.

and

2. $\tau \left| \{\theta_i\} \right.$

The first set of conditionals exploit the fact that the θ_i are conditionally independent. The second set exploit the fact that $\{\theta_i\}$ are sufficient for τ. That is, once the $\{\theta_i\}$ are drawn from 1), these serve as "data" to the inferences regarding τ. If, for example, the first stage prior is normal, then standard natural conjugate priors can be used, and all draws can be done one-for-one and in logical blocks. This normal prior model is also the building block for other more complicated priors. The normal model is given by

$$\theta_i \sim N\left(\bar{\theta}, V_\theta\right)$$

$$\bar{\theta} \sim N\left(\bar{\bar{\theta}}, A^{-1}\right)$$

$$V_\theta \sim IW\left(\upsilon, V\right)$$

In the normal model, the $\{\theta_i\}$ drawn from 1) are treated as a multivariate normal sample and standard conditionally conjugate priors are used. It is worth noting that in many applications the second stage priors are set to be very diffuse ($A^{-1} = 100I$ or larger) and the Wishart is set to have expectation I with very small degrees of freedom such as $\dim(\theta) + 3$. As we often have a larger number of units in the analysis, the data seems to overwhelm these priors and we learn a great deal about τ, or in the case of the normal prior, $\left(\bar{\theta}, V_\theta\right)$.

Drawing the $\{\theta_i\}$ given the unit level data and τ is dependent on the unit level model. For linear models, as illustrated in chapter 3, we can implement a Gibbs sampler by making direct draws from the appropriate conjugate distributions. However, in most marketing applications, there is no convenient conjugate prior or a convenient way of sampling from the conditional posteriors. For this reason, most rely on some sort of

Metropolis algorithm to draw θ_i. As discussed in Chapter 3, there are two very useful variants of the Metropolis algorithm—independence and random walk. Both could be used to develop a general purpose drawing method for hierarchical models. In either case, the candidate draws require a candidate sampling density (as in the case of the independence Metropolis) or an increment density in the case of the RW Metropolis. The performance of these algorithms will depend critically on the selection of these densities.

In both the independence and random walk cases, the densities should be selected to capture the curvature and tail behavior of the conditional posterior

$$p(\theta_i|y_i,\tau) \propto p(y_i|\theta_i)p(\theta_i|\tau) \tag{5.3.1}$$

This suggests that the Metropolis algorithm used to draw each θ_i should be customized for each cross-sectional unit. In the Metropolis literature, there is also a tradition of experimentation with the scaling of the covariance matrix of either the random walk increments or the independence chain candidate density. Clearly, it is not practical to experiment with scaling factors which are customized to each individual unit. In order to develop a practical Metropolis algorithm for hierarchical models, we must provide a method of customization to the unit level which does not require experimentation.

A further complication for a practical implementation is that individual level likelihoods may not have a maximum. For example, suppose the unit level model is a logit model and the unit does not choose all alternatives in the sample. Then there is no maximum likelihood estimator for this unit if we include intercepts for each choice alternative. The unit-level likelihood is increasing in any direction which moves the intercepts for alternatives never chosen to $-\infty$. Most common proposals for Metropolis sampling densities involve use of maximum likelihood estimators. One could argue that the prior in (5.3.1) avoids this problem. Proper priors as well as normal tails will usually insure

that a maximum exists (Allenby and Rossi (1993) suggests an approximate Bayes estimator which uses this posterior mode). However, using the posterior mode (and associated Hessian) in a Metropolis step would require computation of the mode and Hessian at every MCMC step and for each cross-sectional unit[2]. This could render the Metropolis algorithm computationally impractical.

There is a folk-literature on the choices of Metropolis proposal densities for hierarchical models. Some advocate using the same proposal for all units and base this proposal on the asymptotic normal approximation to the pooled likelihood. Obviously, the pooled likelihood is a mixture of unit level likelihoods so that it is possible that this proposal (even if scaled for the relative number of unit-level and total observations) has a location and curvature that is different from any single unit level likelihood. Another popular proposal is to use the current draw of the prior as the basis of a random walk chain. That is, if the first stage prior is normal, we use the current draw of the variance of the θ as the variance of the proposal distribution. This is clearly not a good idea as it does not adapt to the information content available for a specific unit. If all units have very little information, then the prior will dominate and this idea may work acceptably. However, there are units with a moderate to large amount of information, a RW chain will exhibit very high autocorrelation due to rejected draws which come from a prior which is much less tight than the unit level likelihood.

We propose a class of Metropolis algorithms which use candidate sampling distributions which are customized to the unit level likelihoods but are computationally practical in the sense that they do not require order R (the number of MCMC draws) optimizations but only require an initial set of optimizations. These candidate sampling

[2] Note that the parameters of the prior in (5.3.1) will vary from MCMC step to MCMC step.

distributions can either be used as the basis of an independence or RW Metropolis chain. In addition, our proposal does not require that each unit-level likelihood have a maximum. To handle the problem of non-existence of maxima, we use a "fractional" likelihood approach in which we modify the individual level likelihood (but only for the purpose of a Metropolis proposal density $\ell_i(\theta) = p(y_i|\theta)$) by multiplying by a likelihood with a defined maximum.

(5.3.2) $$\ell_i^*(\theta) = \ell_i(\theta)\,\overline{\ell}(\theta)^{\beta}$$

$\overline{\ell}$ can be the pooled likelihood which almost certainly has a maximum. The β weight is designed to scale the pooled likelihood to the appropriate order so that it does not dominate the unit-level likelihood.

(5.3.3) $$\beta = \frac{n_i}{cN}$$

n_i is the number of observations for the ith unit and N is the total number of observations in all units. c is a tuning constant which represents the relative weight of the individual likelihood to the scaled pooled likelihood term. We only bring in the pooled likelihood for the purpose of "regularizing" the problem so we would typically set c at a value greater than 1 (we recommend 2). The pseudo-likelihood[3] in (5.3.2) can be maximized to obtain a location and scale, $\hat{\theta}_i$ and H_i. These quantities can then be combined with the prior to form a Metropolis proposal distribution. In many cases, the prior will be in a normal form so that we can combine the prior and normal approximation to the pseudo unit level likelihood using standard theory for the Bayes linear model. This provides us with a

[3] Computation of the pseudo likelihood estimates need only be performed once prior to initiation of the Metropolis algorithm. It should also be noted that for models without lower dimensional sufficient statistics, the evaluation of the pseudo-likelhood in (5.3.2) requires evaluation of the pooled likelihood as well. To reduce this computational burden we can use the asymptotic normal approximation to the pooled likelihood. As we are using this likelihood only for the purpose of "regularizing" our unit level likelihood, the quality of this approximation is not crucial.

proposal that is customized to the curvature and possible location of each unit-level likelihood. That is, if the prior is $N\left(\bar{\theta}, V_\theta\right)$, then our proposal will be based on a normal density with moments

$$\text{mean: } \theta_i^* = \left(H_i + V_\theta^{-1}\right)^{-1}\left(H_i\hat{\theta}_i + V_\theta^{-1}\bar{\theta}\right)$$

(5.3.4)

$$\text{variance: } \left(H_i + V_\theta^{-1}\right)^{-1} \quad H_i = -\left.\frac{\partial^2 \log\left(\ell_i\right)}{\partial\theta\partial\theta'}\right|_{\theta=\hat{\theta}_i}$$

For an independence Metropolis, we will use both the customized location as well as the curvature estimate for each unit. We note that these will be updated from draw to draw of the prior τ parameters as the chain progresses. However, each update will only use the current draw of τ and the proposal location and scale parameters. A RW chain will use only the scale parameters combined with τ. We will adopt the scaling proposal of Roberts and Rosenthal (2001) and set scaling to $2.93/\sqrt{\dim(\theta)}$. This provides us with two Metropolis algorithms which are automatically tuned. The independence chain might be regarded as somewhat higher risk than the RW chain as we require that both the curvature and location obtained by the approximate pseudo-likelihood procedure be correct. If the location is off, the Independence chain can be highly autocorrelated as it rejects "out-of-place" candidates. The RW chain will adapt to the location of the unit level likelihoods but this could be at the price of higher autocorrelation. Thus, it is the risk-averse alternative.

5.4 A Hierarchical Multinomial Logit Example

To examine the performance of various proposed chains, we consider first the case of a hierarchical logit model. Each of the units is assumed to have an MNL likelihood and we

specifiy a normal distribution of the logit parameters over units with mean $\Delta' z_i$ as in (5.2.1).

The hierarchical logit model takes the form:

$$
\begin{aligned}
&\ell\left(\beta_i \middle| y_i, X_i\right) \ [\text{MNL}] \\
&B = Z\Delta + U \quad u_i \sim N\left(0, V_\beta\right) \\
&\text{vec}\left(\Delta \middle| V_\beta\right) \sim N\left(\text{vec}\left(\bar{\Delta}\right), V_\beta \otimes A^{-1}\right) \\
&V_\beta \sim IW\left(\upsilon, V\right)
\end{aligned}
\tag{5.4.1}
$$

u_i and β_i are the ith rows of B and U. The DAG for the model in (5.4.1) is given by

$$
\begin{array}{ccc}
V_\beta & \searrow & \\
\downarrow & & \beta_i \rightarrow \ y_i \\
\Delta & \nearrow &
\end{array}
\tag{5.4.2}
$$

Given a draw of B, draws of Δ, V_β can be made using standard conjugate theory for the MRM. We can define three possible chains for drawing the β_i:

i) An Independence Metropolis with draws from a multivariate student t with location and scale given by (5.3.4). Note that both the location and scale will be influenced by the current draw of both Δ and V_β. Candidates will be drawn from an Mvst with mean $\beta^* = \left(H_i + \left(V_\beta^r\right)^{-1}\right)^{-1}\left(H_i\hat{\beta}_i + \left(V_\beta^r\right)^{-1}\left(\Delta^r\right)' z_i\right)$ and covariance proportional to $\left(H_i + \left(V_\beta^r\right)^{-1}\right)^{-1}$.

ii) A RW Metropolis with increments having covariance $s^2 V_\beta^r$ where s is a scaling constant and V_β^r is the current draw of V_β.

iii) An improved RW Metropolis with increments having covariance $s^2\left(H_i + \left(V_\beta^r\right)^{-1}\right)^{-1}$ where H_i is the Hessian of the ith unit likelihood evaluated at the MLE for the fraction likelihood defined by multiplying the MNL unit likelihood by the pooled likelihood raised to the β power.

We will choose an "automatic" tuning scheme in which we set $s = 2.93/\sqrt{\dim\left(\beta_i\right)}$.

We might expect the chain defined by i) to perform well if our location estimates for the posteriors of each MNL given the current draw of Δ and V_β are good following the intuition developed for the single logit model in chapter 3. However, we must recognize that the normal approximation to the logit likelihood may break down for likelihoods with no defined maximum as we have if a unit does not choose from all alternatives available. The independence Metropolis chain will not adapt to the proper location unlike the RW chains.

The RW chains offer adaptability in location at the expense of possibly slower navigation. The RW chain defined by ii) which simply uses V_β for the covariance of increments is not expected to perform well for cases in which some units have a good deal of information and others very little. For units with little information, the unit-level conditional posteriors are dominated by the prior term (the unit likelihood is relatively flat) and the RW defined by ii) may have increments of approximately the right scale. However, if a unit has a more sharply defined likelihood, the increments proposed by the chain in ii) will be too large on average. This could create high autocorrelation due to the rejection of candidates and the consequent "stickiness" of the chain. The RW sampler defined by iii) does not suffer from this problem.

To investigate the properties of these three chains, we simulated data from a 5 choice hierarchical logit model with 4 intercepts and one X variable drawn as unif(-1.5, 0) which is meant to approximate a log-price variable with a good deal of variation. 100 units were created: 50 with only 5 observations and 50 with 50 observations. $\beta_i \sim N(\mu, V_\beta)$. $\mu' = (1,-1,0,0,-3)$; V_β has diagonal elements all equal to 3 and with the [4,5] and [5,4] elements set to 1.5. Diffuse priors were used, $A = .01$, $\upsilon = 5+3$, $V = \upsilon I_5$.

Figure 5.1 shows draw sequences (every 20th draw) from the Independence chain (i) and the improved RW chain (iii) for the [5,5] element of V_β. The dark horizontal line is the "true" parameter value. The independence chain takes an extraordinary number of draws to dissipate the initial conditions. It appears to take at least 15,000 draws to reach the stationary distribution. On the other hand, the improved RW chain mixes well and dissipates the initial condition in fewer than 500 iterations. Figure 5.2 compares the two RW chains for a unit with 50 observations. The RW chain with increments based on V_β alone exhibits very poor mixing. Long runs of repeated values are shown in the figure as the chain rejects draws too far away from the mass of the posterior. The improved RW chain shows much better mixing. The numerical efficiency measure (see section 3.9) for the improved RW chain is 3.83 vs 8.66 for the RW chain proposed in ii). For a unit with a small number of observations, both chains have comparable numerical efficiency (5 for the improved chain and 5.94 for the chain proposed in ii)). The improved RW chain is implemented in the *bayesm* function, `rhierMnlRwMixture`.

We also consider an example using scanner panel data on purchases of margarine. This data set was originally analyzed in Allenby and Rossi (1991) and contains purchases on ten brands of margarine and some 500 panelists. This dataset is available in *bayesm* and can be loaded with the command, `data(margarine)`. Several of the brands have only very small share of purchases and, thus, for the purposes of illustration, we consider a subset of data on purchases of six brands: 1). Parkay stick, 2) Blue Bonnett Stick, 3) Fleischmanns stick, 4) House brand stick, 5) Generic stick, and 6) Shed Spread Tub. We also restricted attention to those households with 5 or more purchases. This gives us a data set with 313 households making a total of 3405 purchases. We have brand-specific intercepts and a log price variable in the hierarchy for a total of 6 unit-level logit coefficients. We also have

information on various demographic characteristics of each household including household income and family size which form the Z matrix of observable determinants in the hierarchy. We use "standard" diffuse prior settings of A = .01 I, υ=6+3, V=υI and run the improved RW chain for 20,000 iterations. Figure 5.3 shows the posterior distributions of the price coefficient for selected households. In the top row, the posterior is displayed for two households with a relatively small amount of information. It should be emphasized that these households do not have defined maxima for their unit-level likelihoods. This does not mean that we can't learn something about their price sensitivity. The household level data plus the first-stage prior provide some limited information. In the bottom row of figure 5.3, we display marginal posteriors for households with a larger number of observations. As might be expected, the posteriors sharpen up considerably.

Figure 5.4 shows the marginal posteriors for various functions of V_β. The top histogram shows the marginal posterior of the correlation between the house and generic brand intercepts. This is centered tightly over rather large values, suggesting that household preferences for house and generic brands are highly correlated as has been suggested in the literature on private label brands. In the bottom panel of figure 5.4, the posterior distribution of the standard deviation of the price coefficient is displayed. We note that both quantities are nonlinear functions of V_β; it is inappropriate to apply these functions to the posterior mean, $E\left[V_\beta \,|\text{data} \right]$. Both posterior distributions exhibit substantial skewness and show, yet again, that asymptotic normal approximations to the posterior distribution of key parameters can be poor.

Table 5.1 shows the posterior means (standard deviations) of all correlations in the off-diagonal and the standard deviations of each β on the diagonal. The posterior standard

deviations of the households βs are very large. This shows tremendous heterogeneity between households in brand preference and price sensitivity. Table 5.2 shows that very little of this measured heterogeneity can be attributed to the household demographic attributes, log-income and family size. Most of the elements of Δ displayed in table 5.2 are very imprecisely measured, particularly for the effects of income. Larger families show some preference toward the house and generic brands and shy away from Fleischmanns. However, the general impression is of a weak relationship with these demographic variables.

Figure 5.5 displays the distributions of posterior means of coefficients across the 313 households. These distributions exhibit a good deal of skewness. In section 5.7, we develop a diagnostic for our normal model of household heterogeneity which is based on comparing these distributions to the predictive distribution from our model. The predictive distribution will not be normal as we will integrate out the parameters of the first stage prior. However, for the settings of the hyper-parameters in this data analysis, the predictive distribution will be symmetric, albeit fatter tailed than the normal. This informal evidence suggests that the normal first stage prior may not be adequate. In the next section, we will allow for a more flexible family of priors based on mixtures of normals.

5.5 Using Mixtures of Normals

Much of the work in both marketing and in the general statistics literature has used the normal prior for the first-stage of the hierarchical model. The normal prior offers a great deal of flexibility and fits conveniently with a large Bayesian regression/multivariate analysis literature. The standard normal model can easily handle analysis of many units (Steenburgh, Ainslie and Engebretson 2003), and can include observable determinants of heterogeneity (see Allenby and Ginter 1995; Rossi et al 1996). Typically, we might postulate that various demographic or market characteristics might explain differences in intercepts (brand

preference) or slopes (marketing mix sensitivities). In linear models, this normal prior specification amounts to specifying a set of interactions between the explanatory variables in the model explaining y (see McCulloch and Rossi 1994 for further discussion of this point).

While the normal model is flexible, there are several drawbacks for marketing applications. As discussed above, the thin tails of the normal model tend to shrink outlying units greatly toward the center of the data. While this may be desirable in many applications, it is a drawback in discovering new structure in the data. For example, if the distribution of the unit-level parameters is bi-modal (something to be expected in models with brand intercepts) then a normal first-stage prior may shrink the unit level estimates to such a degree as to mask the multi-modality (see below for further discussions of diagnostics). Fortunately, the normal model provides a building block for a mixture of normals extension of the first-stage prior. Mixtures of normal models provide a great deal of added flexibility. In particular, multiple modes are possible. Fatter tails than the normal can also be accommodated by mixing in normal components with large variance. It is well-known that the mixture of normals model provides a great deal of flexibility and that with enough components, virtually any multivariate density can be approximated. That is to say, we can "build-up" any distribution, no matter how non-normal, using many small normal components in much the same manner as it is possible to build a mountain will small piles of gravel. However, as a practical matter, we may not be able to identify significant deviations from a normal model of heterogeneity as we only observe the unit level parameters with considerable error. Intuition developed by direct application of the mixture of normals approach to estimation of densities for directly observed data may not carry over well to the use of mixture of normals in a hierarchial setting.

The mixture of normals model can also be viewed as a generalization of the popular finite mixture model. The finite mixture model views the prior as a discrete distribution with a set of mass points. This approach has been very popular in marketing due to the interpretation of each mixture point as representing a "segment" and to the ease of estimation. In addition, the finite mixture approach can be given the interpretation of a non-parametric method as in Heckman and Singer (1982). Critics of the finite mixture approach have pointed to the implausibility of the existence of a small number of homogeneous segments as well as the fact that the finite mixture approach does not allow for extreme units whose parameters lie outside the convex hull of the support points. The mixture of normals approach avoids the drawbacks of the finite mixture model while incorporating many of the more desirable features.

The mixture of K multivariate normal models can be used as the basis of the heterogeneity distribution as follows:

$$\begin{aligned} \theta_i &= \Delta' z_i + u_i \\ u_i &\sim N\left(\mu_{ind_i}, \Sigma_{ind_i}\right) \\ ind_i &\sim \text{multinomial}_K\left(\text{pvec}\right) \end{aligned} \tag{5.5.1}$$

ind_i is an indicator latent variable for which component observation i is from. ind takes on values 1, …, K.. pvec is a vector of mixture probabilities of length K. In (5.5.1), the z vector of observable characteristics of the population does not include an intercept and has n_z elements. For this reason, we advise that z be centered so that the mean of θ given average z values will be entirely determined by the normal mixture component means. The moments of θ are given below[4]

[4] The variance can be derived by using the identity $\text{Var}\left(\theta_i\right) = \text{E}\left[\text{Var}\left(\theta_i \middle| \text{ind}\right)\right] + \text{Var}\left(\text{E}\left[\theta_i \middle| \text{ind}\right]\right)$

$$
\begin{aligned}
&\mathrm{E}\left[\theta_i \middle| z_i = \bar{z}, p, \{\mu_k\}\right] = \bar{\mu} = \sum_{k=1}^{K} \mathrm{pvec}_k \mu_k \\
&\mathrm{Var}\left(\theta_i \middle| z_i, p, \{\mu_k\}, \{\Sigma_k\}\right) = \sum_{k=1}^{K} \mathrm{pvec}_k \Sigma_k + \sum_{k=1}^{K} \mathrm{pvec}_k \left(\mu_k - \bar{\mu}\right)\left(\mu_k - \bar{\mu}\right)'
\end{aligned}
\tag{5.5.2}
$$

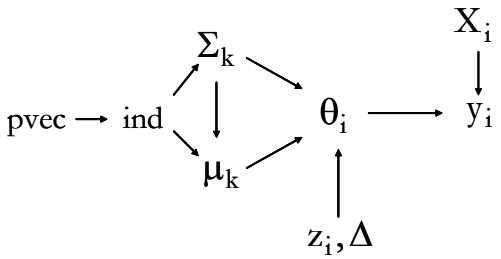
Of course, the variance loses much of its meaning and interpretability as we move farther from an elliptically symmetric distribution.

As in section 3.9, priors for the mixture of normals model can be chosen in convenient conditionally conjugate forms.

$$
\begin{aligned}
&\mathrm{vec}(\Delta) = \delta \sim \mathrm{N}\left(\bar{\delta}, A_\delta^{-1}\right) \\
&\mathrm{pvec} \sim \mathrm{Dirichlet}(\alpha) \\
&\mu_k \sim \mathrm{N}\left(\bar{\mu}, \Sigma_k \otimes a_\mu^{-1}\right) \\
&\Sigma_k \sim \mathrm{IW}(\upsilon, V)
\end{aligned}
\tag{5.5.3}
$$

The DAG for this model can be written as

(5.5.4)



The K parameters in α determine the tightness of the prior on the mixture component probabilities as they are from a natural conjugate prior in which they can be interpreted as cell counts for the components from a previous sample of size, $n^* = \sum_k \alpha_k$. The priors on the mixture components are iid across components in (5.5.3). Much of the statistics literature on mixtures of normals has considered only univariate mixtures of normals or mixture of normals for low dimensional multivariate distributions. Unless there is only a

small amount of data or a very large number of components, the priors for the mixture component parameters may not be very influential in this setting and, thus, may not require too careful consideration. However, in the case of hierarchical models and marketing applications, the mixture model may be applied to parameter vectors of relatively high dimension (such as Multinomial choice model parameters) and the priors will matter as the dimension of parameter space of the normal components may easily exceed 200 or 300.

5.5.1 A Gibbs Sampler

We can easily define an MCMC chain of a Gibbs style by alternating between the draws of individual unit-level parameters and mixture components.

(5.5.5) $$\theta_i \left| \text{ind}_i, \Delta' z_i, \mu_{\text{ind}_i}, \Sigma_{\text{ind}_i} \right.$$

$$\text{pvec}, \text{ind}, \Delta, \{\mu_k\}, \{\Sigma_k\} \left| \{\Theta\} \right.$$

(5.5.6) $$\Theta = \begin{bmatrix} \theta_1' \\ \vdots \\ \theta_m' \end{bmatrix}; \quad \text{ind} = \begin{bmatrix} \text{ind}_1 \\ \vdots \\ \text{ind}_m \end{bmatrix}$$

Some advocate margining out the indicators of the mixture components and using a direct Metropolis step for (5.5.6). The argument here is that removal of these latent variables may improve the mixing of the chain. The likelihood function for the mixture of normals can be evaluated at very low cost. This is certainly possible for mixture of univariate normals in which one could stack up the means and log variances into a vector to be drawn either piecemeal or in one shot by a RW Metropolis. However, in the case of mixtures of multivariate normals, this would require using a Metropolis algorithm to navigate in a very high dimensional space of positive definite matrices. Experience with Metropolis algorithms for covariance structures has shown they are very difficult to tune for satisfactory performance for 5 and higher dimensions. For a mixture of normals, we can easily require a

parameter space with as many as 10 covaraince matrices, each one of which might have 20 or more parameters.

The draw of the hierarchical parameters in (5.5.6) can be broken down into a succession of conditional draws (note the prior parameters are suppressed in (5.5.7) to focus discussion on the nature of the draws).

$$
\begin{aligned}
&\text{ind}\,\big|\,\text{pvec}, Z, \Delta, \{\mu_k, \Sigma_k\}, \Theta \\
&\text{pvec}\,\big|\,\text{ind} \\
&\{\mu_k, \Sigma_k\}\,\big|\,\text{ind}, \Theta \\
&\Delta\,\big|\,\text{ind}, Z, \{\mu_k, \Sigma_k\}, \Theta
\end{aligned}
\tag{5.5.7}
$$

Here we view the $\{\theta_i\}$ or Θ as the "data" generated by a mixture of normals with mean driven by a multivariate regression with explanatory variables in the m x n_z matrix Z. First, we draw the indicators for each component, which provides a classification of the "observations" into one of each of the K components. Given the indicators, there are essentially K independent multivariate normal samples on which conjugate draws can be performed to update the $\{\mu_k, \Sigma_k\}$ parameters. We also must update our views on the mixture probabilities. Since the observable variables in Z affect the means of all components, the draw of Δ must be done by pooling across all observations, adjusting for heteroskedasticity.

The role of the z variables is to shift the mean of the normal mixture on the basis of observations. All of the normal mixture parameters should, therefore, be drawn on the "data" with this component of the mean removed.

$$\Theta^* = \Theta - Z\Delta \tag{5.5.8}$$

As in 3.9, the draw of the indicators is a multinomial draw based on the likelihood ratios with p as the prior probability of membership in each component.

$$\text{ind}_i \sim \text{multinomial}(\pi_i);\ \pi' = (\pi_{i,1},...,\pi_{i,K})$$

(5.5.9)
$$\pi_{i,k} = \text{pvec}_k \frac{\varphi(\theta_i^* | \mu_k, \Sigma_k)}{\sum_k \varphi(\theta_i^* | \mu_k, \Sigma_k)}$$

Here $\varphi(\bullet)$ is the multivariate normal density.

The draw of pvec given the indicators is a Dirichlet draw

(5.5.10)
$$\begin{aligned} &\text{pvec} \sim \text{Dirichlet}(\tilde{\alpha}) \\ &\tilde{\alpha}_k = n_k + \alpha_k \\ &n_k = \sum_{i=1}^{n} I(\text{ind}_i = k) \end{aligned}$$

The draw of each (μ_k, Σ_k) can be made using the algorithm to draw from the Multivariate regression model as detailed in section 2.8.5. For each subgroup of observations, we have an MRM model of the form

(5.5.11)
$$\Theta_k^* = \iota\mu_k' + U;\quad U = \begin{bmatrix} u_1' \\ \vdots \\ u_{n_k}' \end{bmatrix};\ u_i \sim N(0, \Sigma_k)$$

Here Θ_k^* is the submatrix of Θ^* that consists of the n_k rows where $\text{ind}_i = k$. We can use our function, `rmultireg` in ***bayesm***, to achieve these draws.

The draw of Δ requires that we pool data from all K components into one regression model. Since we are proceeding conditional on the component means and variances, we can appropriately standardize the "data" and perform one draw from a standard Bayesian regression model. To motivate the final draw result, let us first consider the kth component. We subset both the Θ and Z matrices to consider only those observations from the kth component and subtract off the mean. Let Θ_k, Z_k be $n_k \times \text{nvar}$ and $n_k \times n_z$ arrays

consisting of only those observations for which $\text{ind}_i = k$. nvar is the dimension of the parameter vectors $\{\theta_i\}$.

(5.5.12) $$Y_k = \Theta_k - \iota\mu_k$$

We can write the model for these observations in the form

(5.5.13) $$Y_k = Z_k\Delta + U_k \quad \text{or} \quad Y_k' = \Delta' Z_k' + U_k'$$

We will stack these nvar equations up to see how to standardize.

(5.5.14) $$\text{vec}\left(Y_k'\right) = \left(Z_k \otimes I_{nvar}\right)\text{vec}\left(\Delta'\right) + \text{vec}\left(U_k'\right)$$

$\text{Var}\left(\text{vec}\left(U_k'\right)\right) = I_{n_k} \otimes \Sigma_k$ and $\Sigma_k = R_k' R_k$. Therefore, if we multiply thru by $I_k \otimes \left(R_k^{-1}\right)'$, this will standardize the error variances in (5.5.14) to have an identity covariance structure.

(5.5.15) $$\left(I_{n_k} + \left(R_k^{-1}\right)'\right)\text{vec}\left(Y_k'\right) = \left(Z_k \otimes \left(R_k^{-1}\right)'\right)\text{vec}\left(\Delta'\right) + z_k$$
$$\text{Var}\left(z_k\right) = I_{n_k \times nvar}$$

We can stack up the K equations of the form of (5.5.15).

(5.5.16)
$$y = X\delta + z$$
$$y = \begin{bmatrix} I_{n_1} \otimes \left(R_1^{-1}\right)' \text{vec}\left(Y_1'\right) \\ \vdots \\ I_{n_K} \otimes \left(R_K^{-1}\right)' \text{vec}\left(Y_K'\right) \end{bmatrix}$$
$$X = \begin{bmatrix} Z_1' \otimes \left(R_1^{-1}\right)' \\ \vdots \\ Z_K' \otimes \left(R_K^{-1}\right)' \end{bmatrix}$$

$\delta = \text{vec}\left(\Delta'\right)$. Given our prior, $\delta \sim N\left(\bar{\delta}, \left(A_\delta\right)^{-1}\right)$, we can combine with (5.5.16) to compute the conditional posterior in the standard normal form.

$$\delta \left| y, X, \bar{\delta}, A_\delta \sim N\left(\left(X'X + A_\delta \right)^{-1} \left(X'y + A_\delta \right), \left(X'X + A_\delta \bar{\delta} \right)^{-1} \right) \right. \tag{5.5.17}$$

The moments needed for (5.5.17) can be calculated efficiently as follows.

$$\begin{aligned} X'X &= \sum_{k=1}^{K} \left(Z_k' Z_k \otimes R_k^{-1} \left(R_k^{-1} \right)' \right) = \sum_{k=1}^{K} \left(Z_k' Z_k \otimes \Sigma_k \right) \\ X'y &= \operatorname{vec}\left(\sum_{k=1}^{K} \Sigma_j^{-1} Y_k' Z_k \right) \end{aligned} \tag{5.5.18}$$

5.5.2 Identification of the Number of Mixture Components

Given that it is possible to undertake posterior simulation of models with 10 or more components, there is some interest in determining the number of components from the data and priors. For mixtures of univariate normals, Richardson and Green (1997) propose an application of the reversible jump sampler that, in principle, allows for MCMC navigation of different size mixture models. The Richardson and Green sampler can "jump" up or down to mixture models of different sizes. In theory, one might be able to use the frequency with which the chain visits a given size component model as an estimate of the posterior probability of that size model. The reversible jump sampler requires a mapping from a lower dimensional mixture component model to a higher dimensional mixture component model. Dellaportas and Papageorgiou (2004) propose a method for extending the Richardson and Green ideas to multivariate mixtures. It remains to be seen how well this will work for the case of a mixture model imbedded within a hierarchical setting.

The other approach to determining the number of mixture components is to attempt to compute the posterior probability of models with a fixed number of components on the basis of simulation output. That is to say, we run 1, 5, and 10 components models and attempt to compute the Bayes Factors for each model. Some have used asymptotic approximations to the Bayes Factors such as the Schwarz approximation. DiCiccio et al

(1997) provide a review of various methods which use simulation output and various asymptotic approximations. All asymptotic methods require finding the posterior mode either by simulation or numerical optimization. This may be particularly challenging in the case of the mixture of multivariate normals in which the likelihood exhibits multiple modes and the parameter space can be extremely high dimensional. In Chapter 6, we will review a number of these methods and return to the problem of computing Bayes Factors for high dimensional models with non-conjugate set-ups. Lenk and DeSarbo (2000) compute Bayes Factors for the number of mixture components in hierarchical generalized linear models.

Given that the normal mixture model is an approximation to some underlying joint density, the main goal in exploring models with different numbers of components is to insure the adequacy of the approximation. That is to say, we want to insure that we include enough components to capture the essential features of the data. The danger of including too many components is that our estimated densities may "overfit" the data. For example, the mixture approximation may build many small lobes on the joint density in an attempt to mimic the empirical distribution of the data in much the same way as kernel smoothing procedures produce lumpy or multimodal density estimates with a too small bandwidth selection. In a hierarchical setting, this is made all the more difficult by the fact that the "data" consist of unknown parameters and we are unable to inspect the empirical distribution. This means that prior views regarding the smoothness of the density are extremely important in obtaining sensible and useful density estimates. The fact that we are in hierarchical setting where the parameters are not observed directly may help us obtain smoother density estimates as the normal mixture will not be able to fit particular noise in the empirical distribution of the "data" as this distribution is only known with error. Thus,

devoting a mixture component to accommodating a few outlying data points will not occur unless these outliers are determined very precisely.

"Testing" for the number of components or computing Bayes Factors is of greater interest to those who wish to attach substantive importance to each component, e..g to proclaim there are X number of subpopulations in the data. It is also possible to use posterior probabilities for model averaging. The purpose of model averaging, in this context, is to insure smoothness by averaging over models of varying numbers of components. Our view is that individual components are not very interpretable and are only included for flexibility. Adequate smoothness can be built in via the priors on mixture component parameters. Thus, we take a more informal approach where we investigate fits from models with varying numbers of components. With informative priors that insure adequate smoothness, addition of components that are not necessary to fit the patterns in the data will not change the fitted density.

5.5.3 Application to Hierarchical Models

The normal mixture model provides a natural generalization to the standard normal model of the distribution of heterogeneity. In this section, we will apply this model to a hierarchical MNL model. In the literature on mixtures of normals, investigators typically use very diffuse informative priors on the mixture component parameters $\{\mu_k, \Sigma_k\}$ and p. Improper priors or even very diffuse proper priors are dangerous in any Bayesian context and especially so in the case of marketing data. Typically, panel data on choice of products includes a subset of panelists who do not purchase from the complete set of products. This means that prior beliefs on the logit model parameters will be very important for this set of panelists. Diffuse but proper priors applied to these panelists will result in the inference that these panelists are essentially never willing to purchase these products under any setting of

the model covariates if product or brand intercepts are included in the model. We will simply set the intercepts for products not purchased to large negative numbers. Given the logistic probability locus, this will result in zero probability of purchase for all intents and purposes. We do not find this plausible. The probabilities of purchase for these products may be low but is not zero. In a one component normal mixture, the other households inform the first stage prior so that we never obtain extreme estimates of intercepts for these panelists with incomplete purchase histories. The thin tails of the normal density as well as reasonable values of the covariance matrix keep us from making extreme inferences. However, in the case of more than one normal component, this can change. If, for example, there are a group of panelists who do not purchase product A, then the mixture model can assign these panelists to one component. Once this assignment is made, the mean product A intercept values for this component will drift off to very large negative numbers. This problem will be particularly acute when a reasonably large number of components are considered.

There are two ways to deal with this problem (note the option of deleting panelists with incomplete purchase records is not defensible!): 1. use models with very small numbers of components or 2. use informative priors on the means of each component. Given that we center the Z variables, the prior on $\{\mu_k\}$ reflects our views regarding intercepts. Recall that we use a $N\left(\bar{\mu}, \Sigma_k \otimes A_\mu^{-1}\right)$ prior. In much of the work with mixtures, this A_μ is set to very small values (e.g. .01). Our view is that this admits implausible intercept values of -20 or +20. This, of course, is only meaningful if all X variables are on the same scale and location. For this reason, we advocate standardizing the X. variables. We then set a value of A_μ of 1/16 or so rather than 1/100.

The prior on Σ_k is also important. If we set a tight prior around a small value, then we may force the normal mixture to use a large number of components to fit the data. In addition, the natural conjugate prior links the location and scale so that tight priors over Σ will influence the range of plausible μ values. We will set the prior on Σ to be relatively diffuse by setting υ to nvar +3 and $V = \upsilon I$.

We return to the margarine example discussed in section 5.4 There are five brand intercepts and one price sensitivity parameter so that nvar=6 and we are fitting a six dimensional distribution of β_i over households. We can combine the Gibbs sampler for normal mixtures with a RW Metropolis chain defined along the lines of 5.3 to draw the household level parameters. We should emphasize that we do have some 300 households but this is not the same as 300 direct observations on six dimensional data. With 300 direct observations, the normal mixture sampler works very well, recovering components with relative ease. However, in the hierarchical logit example, there is only a small amount of information about each household parameter vector. It will be much more difficult to recover complex structure with this effective sample size.

Figure 5.6 presents posterior means of the marginal densities, contrasting one and five component mixture models. For each MCMC draw, we have one fitted multivariate density and we can average these densities over the R draws. To obtain the posterior mean of a marginal density for a specific element of β, we average the marginal densities.

$$\bar{d}_j(\beta_j) = \frac{1}{R}\sum_{k=1}^{K} pvec_k^r \phi_j\left(\beta_j \mid \mu_k^r, \Sigma_k^r\right) \tag{5.5.19}$$

We set down a grid of possible values for each element of β and then evaluate the posterior mean of the density in (5.5.19). These densities are shown in the figure for K=1 and K=5. For at least four of the six elements of β, we see pronounced deviations from normality.

For Fleischmans, House and Generic intercepts, we obtain highly left skewed distributions with some left lobes. Recall that these are intercepts with the base brand set to Parkay stick. This means that these are relative preferences, holding price constant. For Fleischmanns and the House brands there is a mass centered close to zero with a very thick left taill. We can interpret this as that there are a number of households who view the House and Fleischmanns brands as equivalent in quality to Parkay but that there are a number of other households who regard these brands as decidedly inferior.

The non-normality of the estimated first stage prior also has a strong influence on the estimates of household posterior means as illustrated in figure 5.7. There fat tails of the 5 component normal mixture allow for more extreme estimates of both brand intercepts and price sensitivity. Thus, figure 5.7 demonstrates that the observed differences in figure 5.6 make a material difference even if one is only concerned with developing household estimates. However, one should be cautious before using these household estimates. The one component model provides very strong shrinkage of extreme estimates and should, therefore, be regarded as somewhat conservative.

Our fitting strategy for normal mixture models also included adding a large number of components to see if this makes a material difference in the estimated distribution of household level parameters. It should be pointed out that the marginal computation cost of increasing the number of normal components is rather trivial compared to the cost of the Metropolis draws of the individual household parameters. From a computational point of view, the mixture model is basically free and, therefore, can be used routinely. In figure 5.8, we consider a 10 component mixture and compare this to the 5 component. Our view is that the differences in fitted densities in figure 5.8 is rather small. Figure 5.9 compares the

household posterior means for 5 and 10 component models. There is quite close agreement between the estimates derived from 5 and 10 component models.

5.6 Further Elaborations of the Normal Model of Heterogeneity

In many situations, we have prior information on the signs of various coefficients in the base model. For example, price parameters are negative and advertising effects are positive. In a Bayesian approach, this sort of prior information can be included by modifying the first-stage prior. We replace the normal distribution with a distribution with restricted support, corresponding to the appropriate sign restrictions. For example, we can use a log-normal distribution for a parameter which is restricted via sign by the reparameterization, $\theta' = \ln(\theta)$. However, note that this change in the form of the prior can destroy some of the conjugate relationships which are exploited in Gibbs-sampler. However, if Metropolis-style methods are used to generate draws in the Markov chain, it is a simple matter to directly re-parameterize the likelihood function, by substituting exp(θ') for θ, rather than rely on the heterogeneity distribution to impose the range restriction.

What is more important is to ask whether the log-normal prior is appropriate. The left tail of the log-normal distribution declines to zero, insuring a mode for the log-normal distribution at a strictly positive value. For situations in which we want to admit zero as a possible value for the parameter, this prior may not be appropriate. Boatwright et al (1999) explore the use of truncated normal priors as an alternative to the log-normal reparameterization approach. Truncated normal priors are much more flexible allowing for mass to be piled up at zero.

Bayesian models can also accommodate structural heterogeneity, or changes in the likelihood specification for a unit of analysis. The likelihood is specified as a mixture of likelihoods:

$$p(y_{it} \mid \{\theta_{ik}\}) = r_1 p_1(y_{it} \mid \theta_{i1}) + \ldots + r_K p_K(y_{it} \mid \theta_{iK}) \ ,$$

and estimation proceeds by appending indicator variables for the mixture component to the state space. Conditional on the indicator variables, the datum, y_{it}, is assigned to one of K likelihoods. The indicator variables, conditional on all other parameters, have a multinomial distribution with probabilities proportional to the number of observations assigned to the component and the probability that the datum arises from likelihood. Models of structural heterogeneity have been used to investigate intra-individual change in the decision process due to environmental changes (Yang and Allenby 2000) and fatigue (Otter, et.al. 2003).

Finally, Bayesian methods have recently been used to relax the commonly made assumption that the unit parameters, θ_i, are iid draws from the distribution of heterogeneity. Ter Hofstede, Wedel and Steenkamp (2002) employ a conditional Gaussian field specification to study spatial patterns in response coefficients:

$$p(\theta_i \mid \tau) = p(\theta_i \mid \{\theta_j : j \in S_i\}, V_\theta)$$

where S_i denote units that are spatially adjacent to unit i. Since the MCMC estimation algorithm employs full conditional distributions of the model parameters, the draw of θ_i involves using a local average for the mean of the mixing distribution. Yang and Allenby (2003) employ a simultaneous specification of the unit parameters to reflect the possible presence of interdependent effects due to the presence of social and information networks.

$$\theta = \rho W \theta + u$$
$$u \sim N\left(0, \sigma^2 I\right)$$

where W is a matrix that specifies the network, ρ is a coefficient that measures the influence of the network, and u is an innovation. We discuss this model at length in case study 2.

5.7 Diagnostic Checks of the First Stage Prior

In the hierarchical model, the prior is specified in a two stage process:

$$\theta \sim N\left(\bar{\theta}, V_\theta\right)$$
$$p\left(\bar{\theta}, V_\theta\right)$$

In the classical literature, the normal distribution of θ would be called the random effects model and would be considered part of the likelihood rather than part of the prior. Typically, very diffuse priors are used for the second stage. Thus, it is the first stage prior which is important and will always remain important as long as there are only a few observations available per household. Since the parameters of the first stage prior are inferred from the data, the main focus of concern should be on the form of this distribution.

In the econometric literature, the use of parametric distributions of heterogeneity (e.g. normal distributions) are often criticized on the grounds that their mis-specification leads to inconsistent estimates of the common model parameters (c.f. Heckman and Singer (1982)). For example, if the true distribution of household parameters were skewed or bimodal, our inferences based on a symmetric, unimodal normal prior could be misleading. One simple approach would be to plot the distribution of the posterior household means and compare this to the implied normal distribution evaluated at the Bayes estimates of the hyperparameters, $N\left(E\left[\bar{\theta}|\text{data}\right], E\left[V_\theta\right]\right)$. The posterior means are not constrained to follow the normal distribution since the normal distribution is only part of the prior and the posterior is influenced by the unit-level data. This simple approach is in the right spirit but could be misleading since we do not properly account for uncertainty in the unit-level parameter estimates.

Allenby and Rossi (1999) provide a diagnostic check of the assumption of normality in the first stage of the prior distribution that properly accounts for parameter uncertainty. To handle uncertainty in our knowledge of the common parameters of the normal distribution, we compute the predictive distribution of $\theta_{i'}$ for unit i' selected at random from the population of households with the random effects distribution. Using our data and model, we can define the predictive distribution of $\theta_{i'}$ as follows:

$$\theta_{i'} \mid \text{data} = \int\int \phi\left(\theta \mid \bar{\theta}, V_\theta\right) p\left(\bar{\theta}, V_\theta \mid \text{data}\right) d\bar{\theta}\, dV_\theta$$

Here $\phi\left(\theta_{i'} \mid \bar{\theta}, V_\theta\right)$ is the normal prior distribution. We can use our MCMC draws of $\bar{\theta}, V_\theta$, coupled with draws from the normal prior to construct an estimate of this distribution. The diagnostic check is constructed by comparing the distribution of the unit-level posterior means to the predictive distribution based on the model, given above.

5.8 Findings and Influence on Marketing Practice

The last ten years of work on heterogeneity in marketing has yielded several important findings. Researchers have explored a rather large set of first stage models with a normal distribution of heterogeneity across units. In particular, investigators have considered a first stage normal linear regression (Blattberg and George 1991), a first stage logit model (Allenby and Lenk 1994, 1995), a first stage probit (McCulloch and Rossi 1994), a first stage Poisson (Neelamegham and Chintagunta 1999), and a first stage generalized gamma distribution model (Allenby, Leone and Jen (1999)). The major conclusion is that there is a substantial degree of heterogeneity across units in various marketing data sets. This finding of a large degree of heterogeneity holds out substantial promise for the study of preferences, both in terms of substantive and practical significance (Ansari, Jedidi and Jagpal 2000). There may be substantial heterogeneity bias in models that do not properly account

for heterogeneity (Chang, Siddarth and Weinberg 1999), and there is large value in customizing marketing decisions to the unit level (see Rossi, McCulloch and Allenby 1996).

Yang, Allenby and Fennell (2002) investigate the source of brand preference, and find evidence that variation in the consumption environment, and resulting motivations, leads to changes in a unit's preference for a product offering. Motivating conditions are an interesting domain for research as they preexist the marketplace, offering a measure of demand that is independent of marketplace offerings. Other research has documented evidence that the decision process employed by a unit is not necessarily constant throughout a unit's purchase history (Yang and Allenby 2000). This evidence indicates that the appropriate unit of analysis for marketing is at the level that is less aggregate than a person or respondent, although there is evidence that household sensitivity to marketing variables (Ainslie and Rossi 1998) and state dependence (Seetharaman, Ainslie and Chintagunta 1999) is constant across categories.

The normal continuous model of heterogeneity appears to do reasonably well in characterizing this heterogeneity but there has not yet been sufficient experimentation with alternative models such as the mixture of normals to draw any definitive conclusions. With the relatively short panels typically found in marketing applications, it may be difficult to identify much more detailed structure beyond that afforded by the normal model. In addition, relatively short panels may produce a confounding of the finding of heterogeneity with various model mis-specifications in the first stage. If only one observation is available for each unit, then the probability model for the unit level is the mixture of the first stage model with the second stage prior:

$$p(y|\tau) = \int p(y|\theta)p(\theta|\tau)\,d\theta$$

This mixing can provide a more flexible probability model. In the one observation situation, we can never determine whether it is "heterogeneity" or lack of flexibility that causes the Bayesian hierarchical model to fit the data well. Obviously, with more than one observation per unit, this changes and it is possible to separately diagnose first stage model problems and deficiencies in the assumed heterogeneity distribution. However, with short panels there is unlikely to be a clean separation between these problems and it may be the case that some of the heterogeneity detected in marketing data is really due to lack of flexibility in the base model.

There have been some comparisons of the normal continuous model with the discrete approximation approach of a finite mixture model. It is our view that it is conceptually inappropriate to view any population of units of being comprised of only a small number of homogeneous groups and, therefore, the appropriate interpretation of the finite mixture approach is an approximation method. Allenby and Rossi (1999) and Lenk et al (1996) show some of the shortcomings of the finite mixture model and provide some evidence that the finite mixture model does not recover reasonable unit level parameter estimates.

At the same time that the Bayesian work in the academic literature has shown the ability to produce unit-level estimates, there has been increased interest on the part of practitioners in unit-level analysis. Conjoint researchers have always had an interest in respondent-level part-worths and had various ad hoc schemes for producing these estimates. Recently, the Bayesian hierarchical approach to the logit model has been implemented in the popular Sawtooth conjoint software. Experience with this software and simulation studies have led Rich Johnson, Sawtooth software's founder, to conclude that Bayesian methods are superior to others considered in the conjoint literature (Sawtooth Software, 2001).

Retailers are amassing volumes of store-level scanner data. Not normally available to academic researchers, this store-level data is potentially useful for informing the basic retail decisions such as pricing and merchandizing. Attempts to develop reliable models for pricing and promotion have been frustrated by the inability to produce reliable promotion and price response parameters. Thus, the promise of store-level pricing has gone unrealized. Recently, a number of firms have appeared in this space, offering data-based pricing and promotion services to retail customers. At the heart of some of these firms' approach is a Bayesian shrinkage model applied to store-sku-week data obtained directly from the retail client. The Bayesian shrinkage methods produce reasonable and relatively stable store-level parameter estimates. This approach builds directly on the work of Montgomery (1997).

Figure 5.1

Comparison of Draw Sequences: Independence and RW Chains

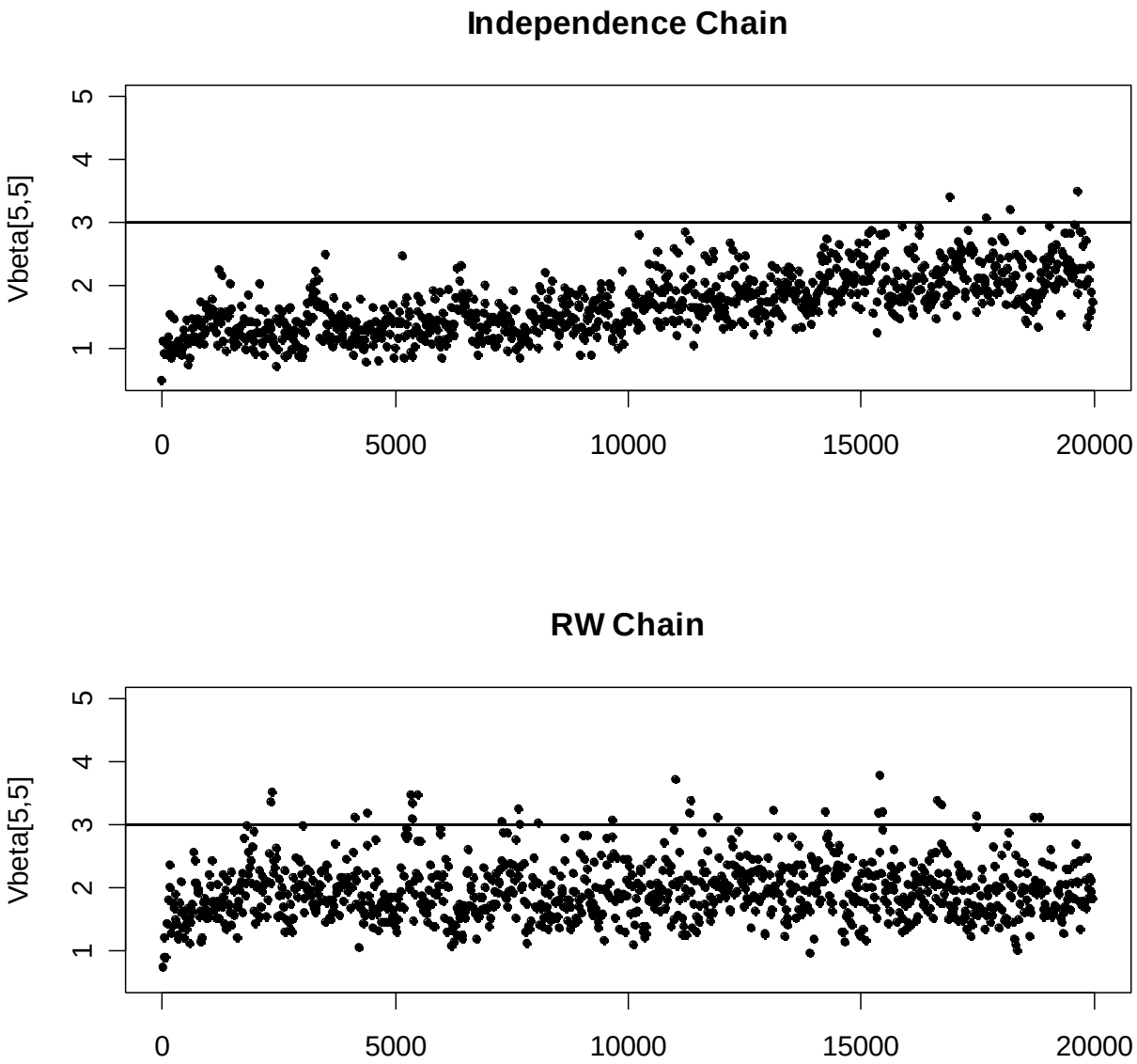


Figure 5.2
Comparison of Two RW Chains:
Draws of Unit Level Parameters

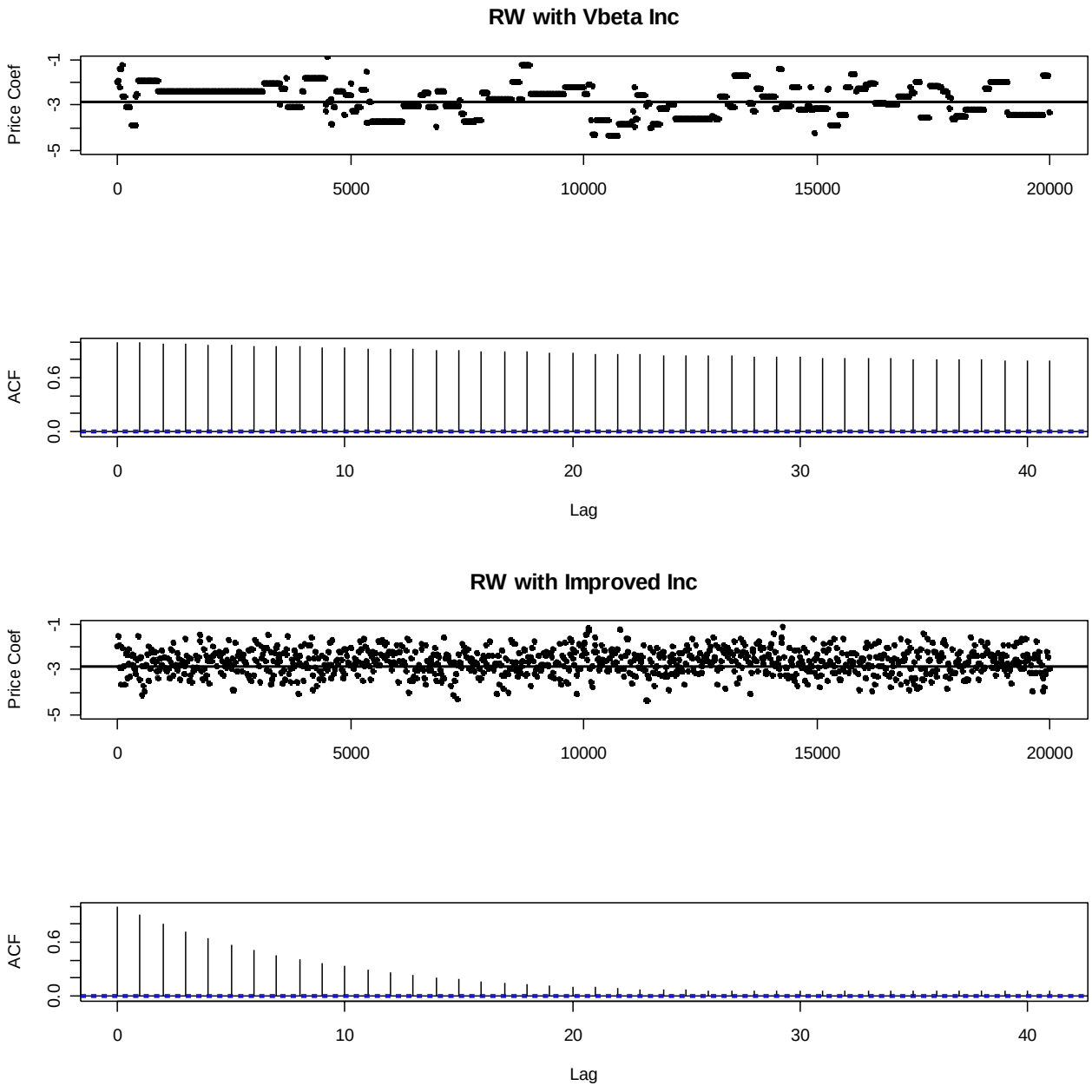


Figure 5.3
Posterior Distribution of Price Coefficients:
Selected Households with Small and Large Number of Observations

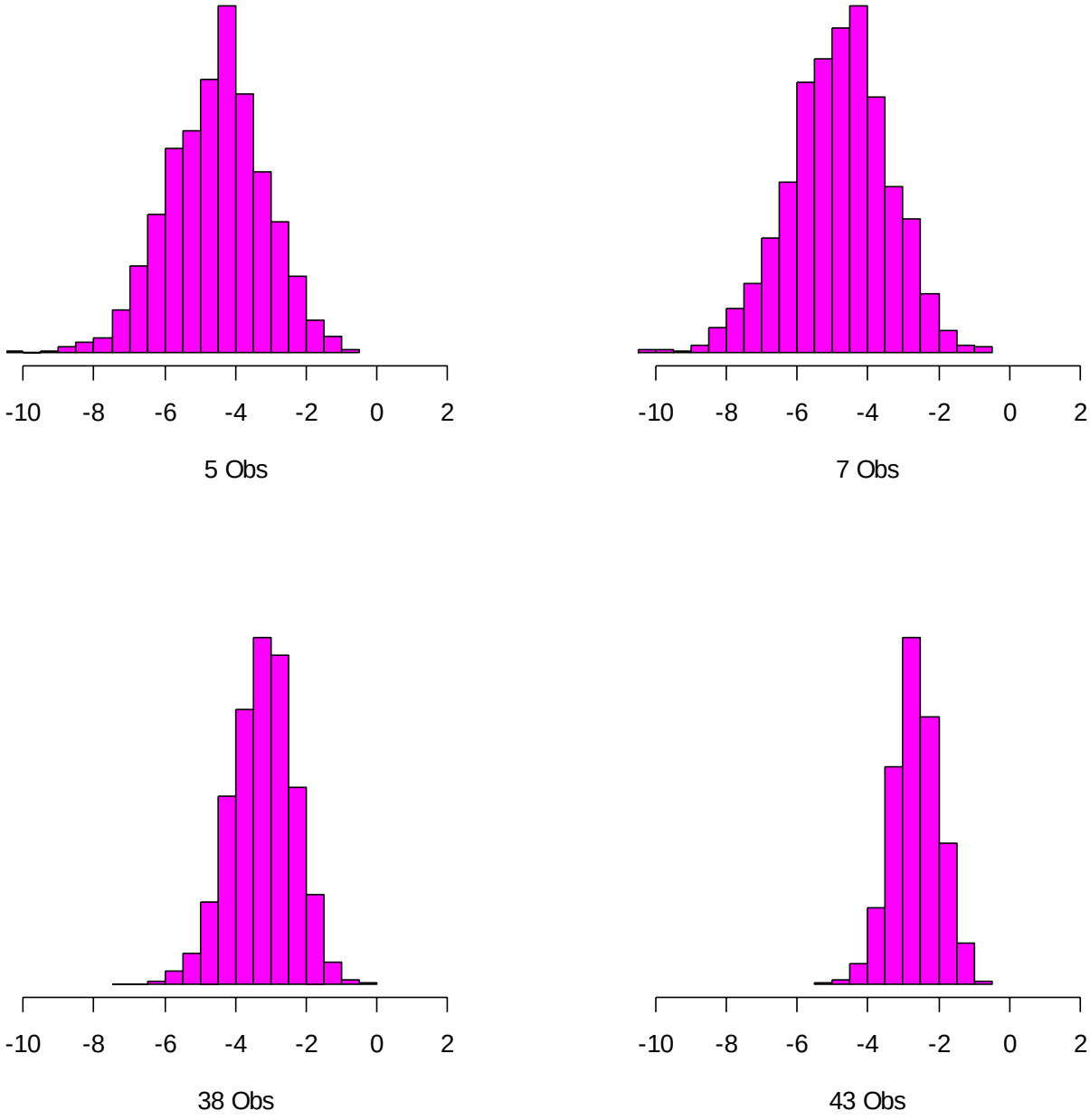
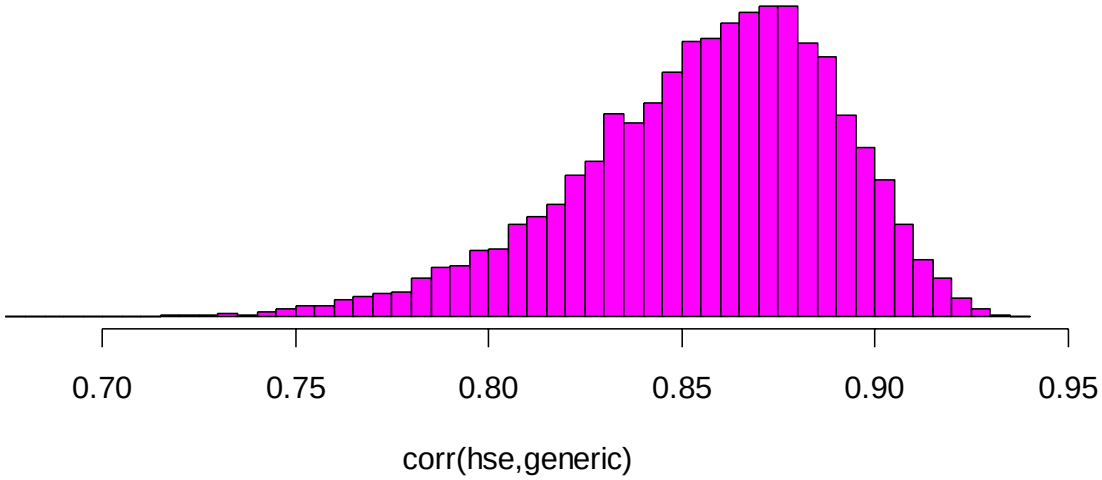


Figure 5.4
Posterior Distribution of Selected Variance Parameters:
Hierarchical Logit Example



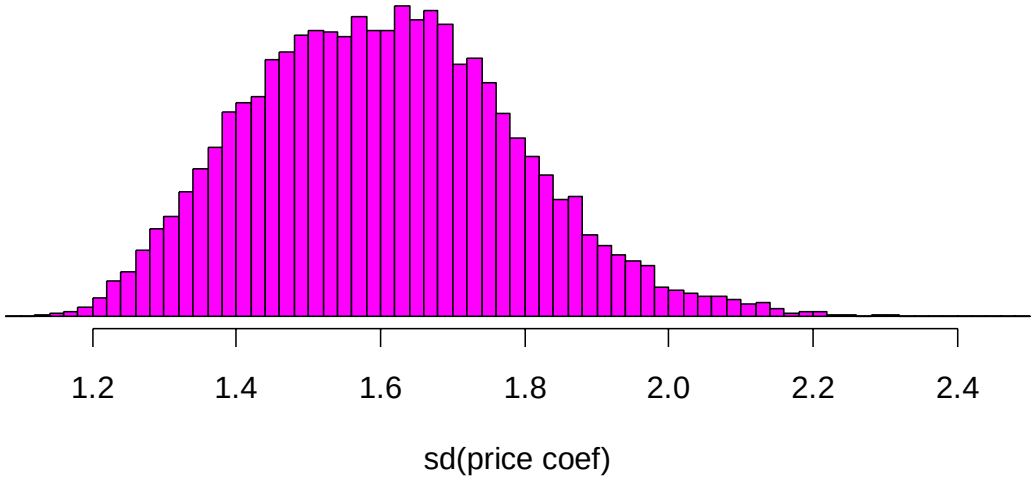


Figure 5.5
Distribution of Posterior Means of Household Coefficients

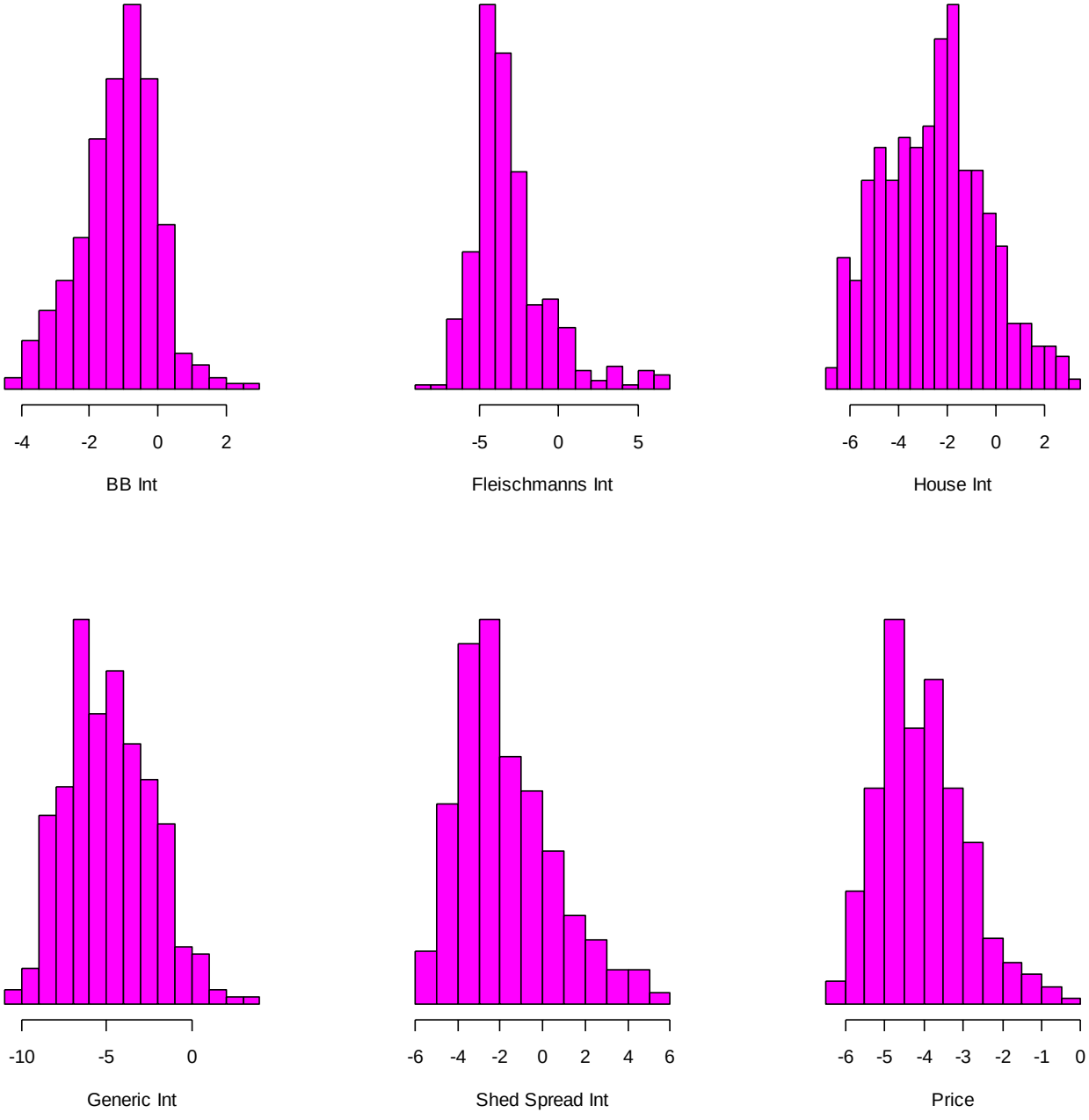


Table 5.1
Correlations and Standard Deviations of Betas:
Hierarchical Logit Example

	BlueBonnet	Fleischmanns	House	Generic	Shed Spred Tub	Price
BlueBonnett	**1.53** (.13)	**.39** (.13)	**.43** (.10)	**.46** (.10)	**.27** (.13)	**-.07** (.14)
Fleischmanns		**3.44** (.55)	**.31** (.15)	**.28** (.18)	**.09** (.17)	**.48** (.15)
House			**2.5** (.19)	**.86** (.03)	**.49** (.10)	**-.05** (.14)
Generic				**3.0** (.27)	**.55** (.10)	**-.08** (.14)
Shed Spread Tub					**3.0** (.33)	**.05** (.15)
Price						**1.6** (.18)

Diagonal contains standard deviations; off-diagonal the correlations.

Table 5.2
Posterior Distribution of Δ:
Hierarchical Logit Example

	BlueBonnet Intercept	Fleischmanns Intercept	House Intercept	Generic Intercept	Shed Spread Intercept	Log(Price)
Intercept	**-1.27** (.64)	**-3.37** (1.8)	**-3.31** (.99)	**-4.96** (1.2)	**.03** (1.2)	**-3.48** (.85)
Log(Income)	**.07** (.21)	**.80** (.59)	**.02** (.32)	**-.51** (.40)	**-.62** (.42)	**-.26** (.28)
Family Size	**-.03** (.10)	**-.70** (.28)	**.24** (.14)	**.55** (.18)	**.06** (.20)	**.08** (.12)

Figure 5.6
1 Vs. 5 Component Mixtures

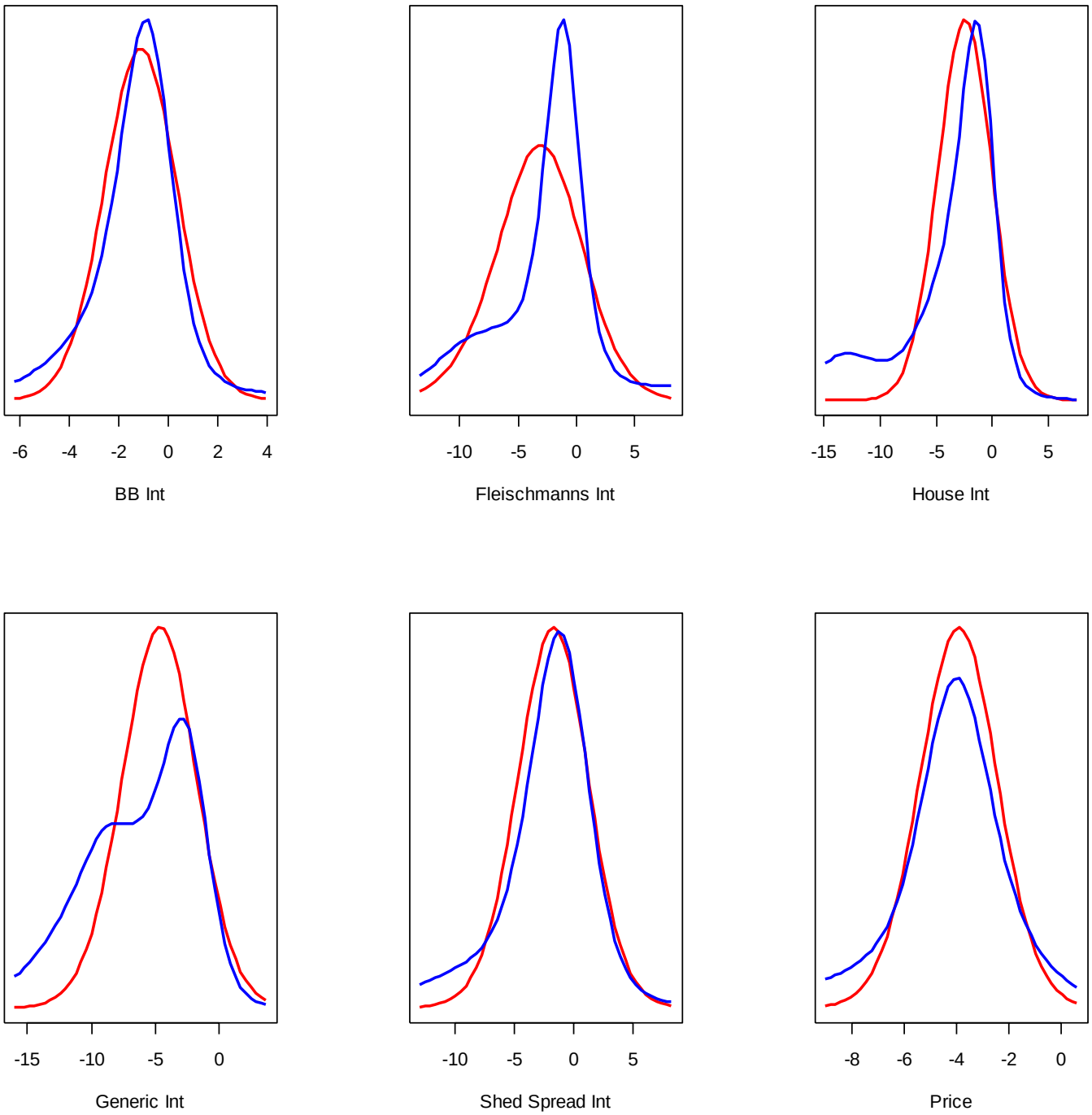


Figure 5.7
Household Coefficient Posterior Means:
1 Vs. 5 Components

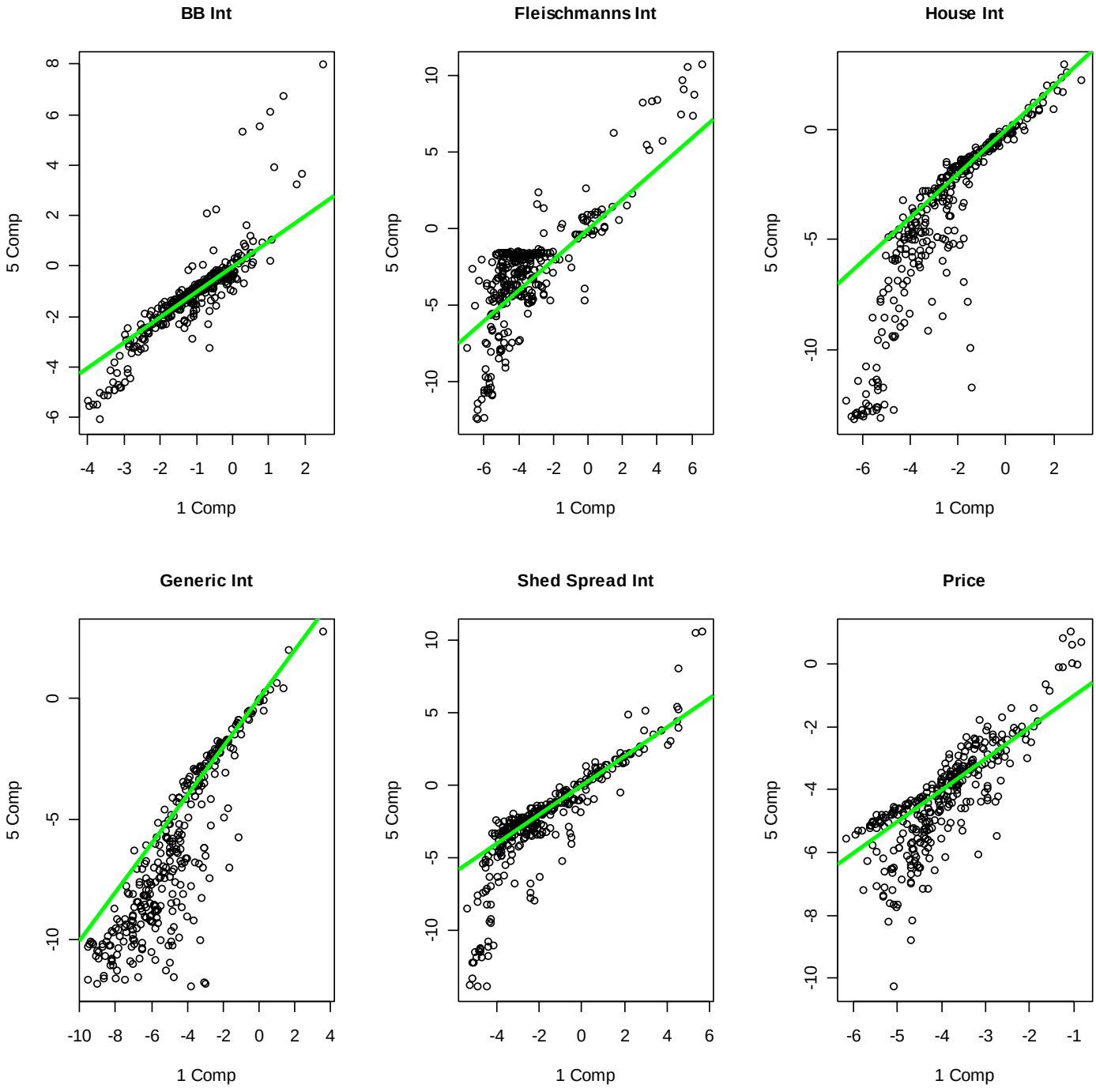


Figure 5.8
5 Vs. 10 Component Mixtures (10 in Red)

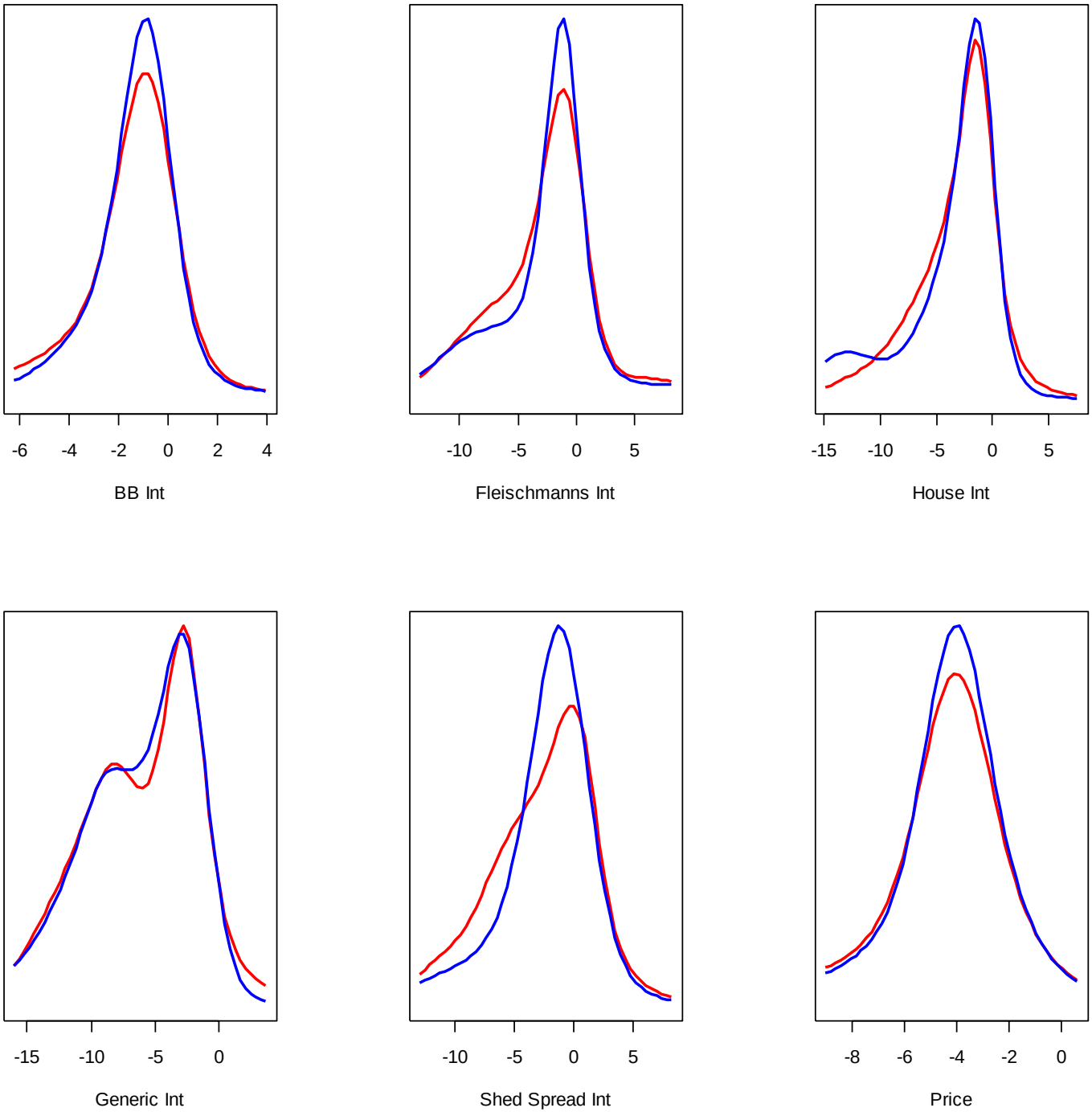
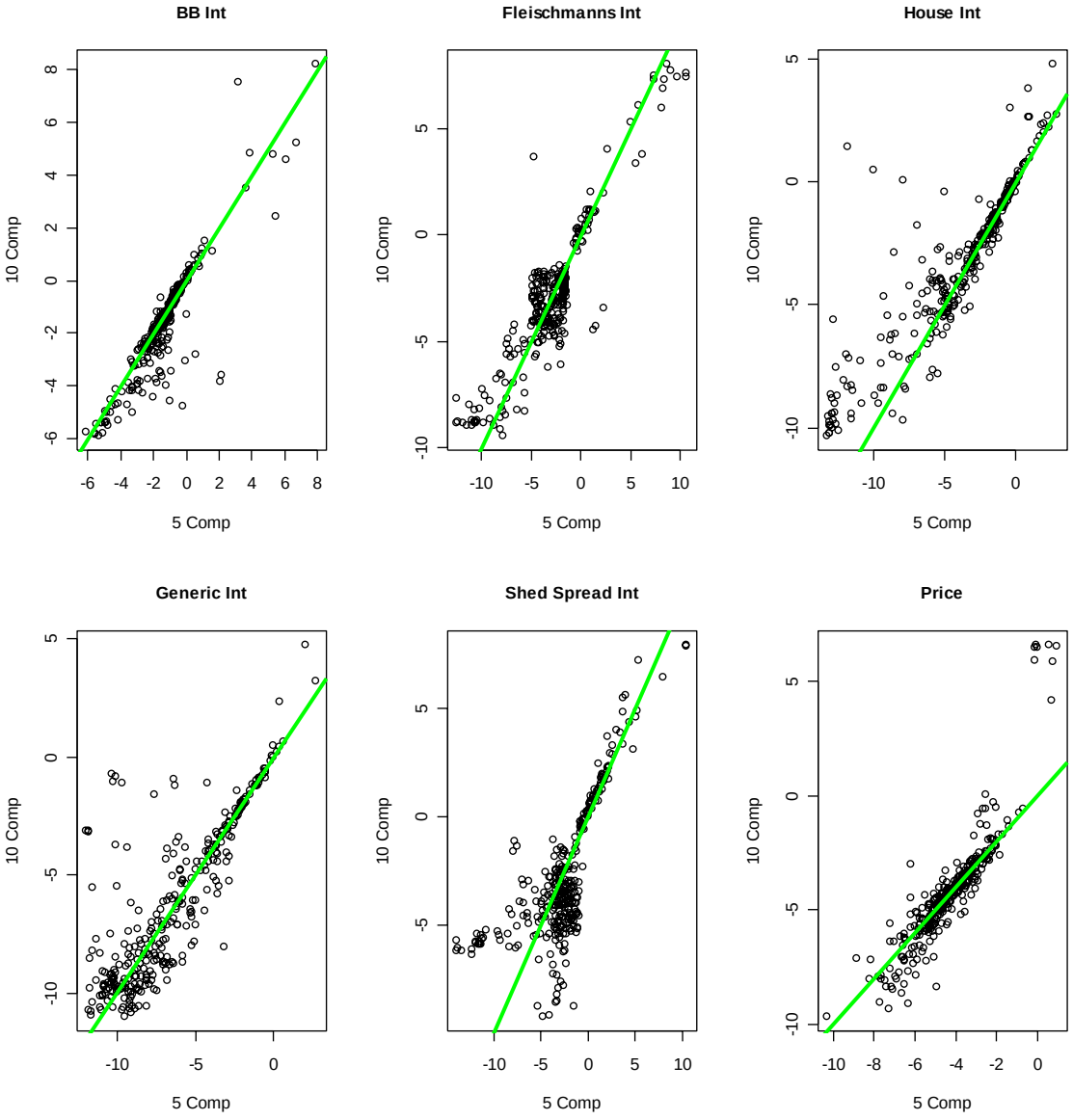


Figure 5.9


BB Int
Fleischmanns Int
House Int
Generic Int
Shed Spread Int
Price
10 Comp
5 Comp