Chapter 6 Model Choice and Decision Theory

Using this Chapter

This chapter discusses Bayesian Model Choice and decision theory. Bayesian Model Choice involves various approaches to computing the posterior probability of a model. Posterior model probabilities are useful in comparing two or more competing models or in the choice from a class of models. While there are some methods which can use standard MCMC output to approximate these probabilities, most problems require additional computations for accurate evaluation of posterior probabilities. Sections 6.1-6.9 introduce various methods for computation of model probabilities and compare some of the most useful methods in the context of a model comparison motivated by the multinomial probit model. Many marketing problems suggest a natural decision problem (such as profit-maximization) so that there is more interest in non-trivial applications of decision theory. Sections 6.10 and 6.11 introduce a Bayesian decision-theoretic approach to marketing problems and provide an example by considering the valuation of disaggregate sample information.

Most of the recent Bayesian literature in marketing emphasizes the value of the Bayesian approach to inference, particularly in situations with limited information. Bayesian inference is only a special case of the more general Bayesian decision-theoretic approach. Bayesian Decision Theory has two critical and separate components: 1. a loss function and 2. the posterior distribution. The loss function associates a loss with a state of nature and an action, $\ell(a,\theta)$ where *a* is the action and θ is the state of nature (parameter). The optimal decision maker chooses the action so as to minimize expected loss where the expectation is taken with respect to the posterior distribution.

$$\min_a \overline{\ell}(a) = \int \ell(a,\theta)\, p(\theta|\text{Data})\, d\theta$$

As indicated in chapter 2, inference about θ can be viewed as a special case of decision theory where the "action" is to choose an estimate based on the data. Quadratic loss yields the posterior mean as the optimal (minimum expected loss) estimator.

Model choice can also be thought of as a special case of decision theory where there is a zero-one loss function. If loss is "1" when the correct model is chosen and "0" if not, then the optimal action is to choose the model with highest posterior probability. In sections 6.1-6.9, we develop methods for computing posterior probabilities for a set of models.

In marketing applications, a natural loss function is the profits of a firm. The firm seeks to determine marketing actions so as to maximize the expected profits which arise from these actions. In section 6.10, we develop this loss frame work and apply this idea to the valuation of various information sets. An example of targeting couponing is introduced to make these ideas concrete in section 6.11.

6.1 Model Selection

In many scientific settings, the action is a choice between competing models. In the Bayesian approach, it is possible to define a set of models $M_1,\ldots,M_k$ and calculate the posterior probability of each of the models. If the loss function is zero when the correct model is chosen and equal to one for all cases in which the incorrect model is chosen, then the optimal Bayesian decision maker chooses the model with the highest posterior probability. The posterior probability of a model can be obtained from the data likelihood and the prior model probability in the usual manner. Throughout this chapter, we will use the notation, "y", to refer to the observed data. This is for ease of reference to the Bayesian model choice literature.

$$p(M_i|y) = \frac{p(y|M_i)p(M_i)}{p(y)} \tag{6.1.1}$$

If the set of models, $\{M_1,\ldots,M_k\}$ is exhaustive, we can compute the posterior probability of model i as

(6.1.2) $$p(M_i|y) = \frac{p(y|M_i)p(M_i)}{\sum_j p(y|M_j)p(M_j)}$$

In many instances, we might wish to compare two models and choose the model with higher posterior probability. For these cases, the ratio of the posterior probabilities (called the posterior odds ratio) is relevant. The posterior odds ratio is the ratio of relative likelihood times the prior odds ratio.

(6.1.3) $$\frac{p(M_1|y)}{p(M_2|y)} = \frac{p(y|M_1)}{p(y|M_2)} \times \frac{p(M_1)}{p(M_2)}$$
$$= \text{Bayes Factor} \times \text{Prior Odds}$$

In a parametric setting, the posterior probability of a model requires that we integrate out the parameters by averaging the density of the data over the prior density of the parameters,

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

Some write this as the expectation of the likelihood with respect to the prior distribution and, thus, call this the "marginal likelihood."

(6.1.5) $$\ell^*(y|M_i) = E_{\theta|M_i}\left[\ell(\theta|y, M_i)\right]$$

However, it should be noted that the likelihood is any function proportional to the data density so that this interpretation is somewhat imprecise. Henceforth in this chapter, when we use the notation $\ell(\bullet)$, we mean the density of the data including all normalizing constants. The intuition is that if the likelihood of the model is high where we think the parameters are (a priori) then the model has high posterior probability.

The Posterior Odds ratio for parametric models can be written:

$$(6.1.6)\quad \frac{p(M_1|y)}{p(M_2|y)} = \frac{\int \ell_1(\theta_1)p_1(\theta_1)d\theta_1}{\int \ell_2(\theta_2)p_2(\theta_2)d\theta_2} \times \frac{p(M_1)}{p(M_2)} = BF \times \text{Prior Odds}$$

In the Bayesian approach, the posterior probability only requires specification of the class of models and the priors. There is no distinction between nested and non-nested models as in the classical hypothesis-testing literature. However, we do require specification of the class of models under consideration; there is no omni-bus measure of the plausibility of a given model or group of models versus some un-specified and possibly unknown set of alternative models.

In the classical testing literature, there is an important distinction made between the non-nested and nested cases. The Neyman-Pearson approach to hypothesis-testing requires the specification of a specific null hypothesis. In the case of non-nested models, there is no natural "null" model and classical methods for hypothesis-testing can lead to contradictory results in which model 1 is rejected in favor of model 2 and vice versa. The Bayesian approach uses the predictive density of the data under each model in the Bayes Factor (6.1.4). The predictive density is the density of the data averaged over the prior distribution of the model parameters. This can be defined for any set of models, nested or non-nested.

(6.1.4) also reveals the sensitivity of the Bayes Factor to the prior. This is not a limitation or weakness of the Bayesian approach but simply a recognition that model comparison depends critically on the assessment of a prior. From a practical point of view, this means that the researcher must carefully select the prior for each model to insure that it does indeed reflect his views regarding the possible set of parameter values. In particular, "standard" diffuse prior settings can be deceptive. As the diffusion of the prior increases for a particular model, the value of the predictive density in (6.1.4) will decline, at least in the limit. This means that relative diffusion of the priors is important in the computation of the

Bayes Factor for comparison of two models. If one prior is very diffuse relative to the other, the BF will tend to favor the model with the less diffuse prior. Improper priors are the limiting case of proper, but diffuse priors, and must be avoided in the computation of Bayes Factors.

We also can interpret the marginal density of the data given model i as the normalizing constant of the posterior.

$$p\left(\theta|y,M_i\right)=\frac{\ell\left(\theta|y,M_i\right)p\left(\theta|M_i\right)}{p\left(y|M_i\right)} \tag{6.1.7}$$

If we let $\tilde{p}\left(\theta|y,M_i\right)$ denote the "un-normalized" posterior, then the marginal density of the data can be written:

$$p\left(y|M_i\right)=\int\tilde{p}\left(\theta|y,M_i\right)d\theta=\frac{\tilde{p}\left(\theta|y,M_i\right)}{p\left(\theta|y,M_i\right)} \tag{6.1.8}$$

6.2 Bayes Factors in the Conjugate Setting

For conjugate models, (6.1.8) can be used to compute Bayes Factors. That is, the full density form for the posterior can be divided by the product of the data density and the prior to obtain the marginal likelihood of the data. This requires, however, fully conjugate priors not just conditionally conjugate priors. Care must be taken to include all appropriate normalizing constants of the posterior, data density and priors.

$$p\left(y|M_i\right)=\frac{\tilde{p}\left(\theta|y,M_i\right)}{p\left(\theta|y,M_i\right)}=\frac{p\left(y|\theta,M_i\right)p\left(\theta|M_i\right)}{p\left(\theta|y,M_i\right)}$$

For nested hypotheses, a simplification of the Bayes Factor, called the Savage-Dickey density ratio, can be used. Consider the case of comparison of model M_0 and M_1 where

M_0 is a restricted version of M_1. Transform[1] θ to ϕ so that the restriction amounts to setting a subvector of ϕ to some specified value (often 0).

$M_0: \ \phi_1 = \phi_1^h$
$M_1:$ unrestricted

where $\phi' = \left(\phi_1', \phi_2'\right)$. In this case, the Bayes Factor for comparison of M_0 to M_1 is given by

$$\frac{p\left(y|M_0\right)}{p\left(y|M_1\right)} = \frac{\int \ell\left(\phi_2|y\right) p\left(\phi_2\right) d\phi_2}{\iint \ell\left(\phi_1,\phi_2|y\right) p\left(\phi_1,\phi_2\right) d\phi_1\, d\phi_2} \tag{6.2.1}$$

where $\ell\left(\phi_2|y\right) = \ell\left(\phi_1,\phi_2|y\right)\Big|_{\phi_1=\phi_1^h}$.

One "natural" choice for the prior on the unrestricted component, ϕ_2, is the conditional prior derived from the joint prior under M_1.

$$p\left(\phi_2\middle|\phi_1 = \phi_1^h\right) = \frac{p\left(\phi_1,\phi_2\right)}{\int p\left(\phi_1,\phi_2\right) d\phi_2}\Bigg|_{\phi_1=\phi_1^h} \tag{6.2.2}$$

Using the prior in (6.2.2), the BF can be written

$$\begin{aligned} BF &= \frac{\int \ell\left(\phi_1,\phi_2|y\right) \frac{p\left(\phi_1,\phi_2\right)}{p\left(\phi_1\right)} d\phi_2\Big|_{\phi_1=\phi_1^h}}{\iint \ell\left(\phi_1,\phi_2\right) p\left(\phi_1,\phi_2\right) d\phi_1\, d\phi_2} \\ &= \frac{\int \ell\left(\phi_1,\phi_2|y\right) p\left(\phi_1,\phi_2\right) d\phi_2}{p\left(\phi_1\right) \iint \ell\left(\phi_1,\phi_2|y\right) p\left(\phi_1,\phi_2\right) d\phi_1\, d\phi_2}\Bigg|_{\phi_1=\phi_1^h} \\ &= \frac{\int p\left(\phi_1,\phi_2|y\right) d\phi_2}{p\left(\phi_1\right)}\Bigg|_{\phi_1=\phi_1^h} \end{aligned} \tag{6.2.3}$$

[1] In most cases, this must be a linear transformation to take full advantage of the Savage-Dickey simplication.

Thus, the BF can be written as the ratio of the marginal posterior of ϕ_1 to the marginal prior of ϕ_1 evaluated at $\phi_1 = \phi_1^h$. We also note that if the conditional posterior of $\phi_1|\phi_2$ is of known form, then we can write the marginal posterior as the average of the conditional. We can estimate this with MCMC output from the marginal posterior of ϕ_2.

$$\text{BF} = \left. \frac{\int p(\phi_1|\phi_2, y)\, p(\phi_2|y)\, d\phi_2}{p(\phi_1)} \right|_{\phi_1=\phi_1^h}$$

(6.2.4)

$$\hat{\text{BF}} = \left. \frac{\frac{1}{R}\sum_r p(\phi_1|\phi_2^r, y)}{p(\phi_1)} \right|_{\phi_1=\phi_1^h}$$

Outside of the conjugate setting, the computation of Bayes Factors must rely on various numerical methods for computing the requisite integrals unless asymptotic methods are used. For most problems, asymptotic methods which rely on the approximate normality of the posterior are not reliable.[2] We briefly review the asymptotic approach to computation of Bayes Factors.

6.3 Asymptotic Methods for Computing Bayes Factors

Asymptotic methods can be used to approximate model probabilities. The idea is that the posterior converges to a normal distribution and then we can use results for the multivariate normal distribution to approximate the marginal likelihood and, therefore, the posterior model probability.

(6.3.1) $$p(y|M_i) = \int p(y|\theta_i, M_i)\, p(\theta_i|M_i)\, d\theta_i$$

We can approximate the integral in (6.3.1) using the normal approximation to the posterior. This is achieved by expanding the log of the un-normalized posterior around its mode in a Taylor series.

$$
\begin{aligned}
p\left(y\middle|M_i\right) &= \int \exp\left(\Gamma(\theta)\right)d\theta \\
&\approx \int \exp\left(\Gamma\left(\tilde{\theta}\right) - \tfrac{1}{2}\left(\theta-\tilde{\theta}\right)' H\left(\tilde{\theta}\right)(\bullet)\right)d\theta = \exp\left(\Gamma\left(\tilde{\theta}\right)\right)(2\pi)^{p/2}\left|H\left(\tilde{\theta}\right)\right|^{-1/2} \\
&= p\left(y\middle|\tilde{\theta}, M_i\right)p\left(\tilde{\theta}\middle|M_i\right)(2\pi)^{p/2}\left|H\left(\tilde{\theta}\right)\right|^{-1/2}
\end{aligned} \tag{6.3.2}
$$

$\tilde{p}\left(\theta|y\right) = \exp\left(\Gamma(\theta)\right)$. $\tilde{\theta}$ is the posterior mode and $H\left(\tilde{\theta}\right) = -\left.\dfrac{\partial^2 \tilde{p}\left(\theta|y\right)}{\partial\theta\partial\theta'}\right|_{\theta=\tilde{\theta}}$, the negative of the Hessian of the un-normalized posterior. The approximate BF for comparison of two models will depend on the ratio of the likelihoods as well as the ratio of the prior densities evaluated at the posterior mode.

$$
BF = \frac{p\left(y\middle|M_1\right)}{p\left(y\middle|M_2\right)} \approx \frac{p\left(\tilde{\theta}_1\middle|M_1\right)}{p\left(\tilde{\theta}_2\middle|M_2\right)} \times \frac{p\left(y\middle|\tilde{\theta}_1, M_1\right)\left|H_1\left(\tilde{\theta}_1\right)\right|^{-1/2}}{p\left(y\middle|\tilde{\theta}_2, M_2\right)\left|H_2\left(\tilde{\theta}_2\right)\right|^{-1/2}} \times (2\pi)^{(p_1-p_2)/2} \tag{6.3.3}
$$

If we make one of the priors in (6.3.3) more diffuse, then the prior density evaluated at the posterior mode will decline and the approximate BF will move in favor of the other model.

Of course, it is possible to base the asymptotic approximation by expanding about the MLE rather than the posterior mode.

$$
BF = \frac{p\left(y\middle|M_1\right)}{p\left(y\middle|M_2\right)} \approx \frac{p\left(\hat{\theta}_{MLE,1}\middle|M_1\right)}{p\left(\hat{\theta}_{MLE,2}\middle|M_2\right)} \times \frac{p\left(y\middle|\hat{\theta}_{MLE,1}, M_1\right)\left|Inf_1\left(\hat{\theta}_{MLE,1}\right)\right|^{-1/2}}{p\left(y\middle|\hat{\theta}_{MLE,2}, M_2\right)\left|Inf_2\left(\hat{\theta}_{MLE,2}\right)\right|^{-1/2}} \times (2\pi)^{\frac{p_1-p_2}{2}} \tag{6.3.4}
$$

$Inf_i\left(\theta\right) = \left[\dfrac{\partial^2 \log \ell_i}{\partial\theta\partial\theta'}\right]$ is the observed information matrix. Note that this is the observed information in the sample of size N. For vague priors, there will be little difference. However, for informative priors, maximum accuracy can be obtained by expanding about the posterior mode. Computation of the posterior mode is no more difficult or time

[2] One notable exception is the MNL model whose likelihood, we have already seen, closely resembles a normal

intensive than computation of the MLE. Moreover, the posterior may be a more regular surface to maximize over than the likelihood. An extreme example of this occurs when the maximum of the likelihood fails to exist or in cases of non-identified parameters.

The approximate BFs in (6.3.3) and (6.3.4) differ only by constants which do not depend on n. While these two approximations are asymptotically equivalent, we have reason to believe that the BF based on the posterior mode may be more accurate. Both expressions are dependent on the ordinate of the prior, while the expression in (6.3.3) also depends on the curvature of the prior (the curvature of the log posterior is the sum of the curvature of the prior and the likelihood). It is possible to define an asymptotic approximation to the marginal density of the data which depends only on the dimension of the model (see Schwarz (1978)). If we expand about the MLE and rewrite the information matrix in terms of the average information, the posterior probability of model i can be written as

$$\begin{aligned} p\left(M_i|y\right) &\propto p\left(y|M_i\right) \\ &\approx k_i p\left(\hat{\theta}_{MLE,i}|M_i\right) p\left(y|\hat{\theta}_{MLE,i}, M_i\right)(2\pi)^{\frac{p_i}{2}} n^{-\frac{p_i}{2}} \left|Inf_i\left(\hat{\theta}_{MLE,i}\right)/n\right|^{-\frac{1}{2}} \end{aligned} \tag{6.3.5}$$

p_i is the dimension of the parameter space for model i. Asymptotically, the average information converges to the expected information in one observation. If we drop everything that is not of order n, then (6.3.5) simplifies to

$$p\left(M_i|y\right) \approx p\left(y|\hat{\theta}_{MLE,i}, M_i\right) n^{-p_i/2} \tag{6.3.6}$$

(6.3.6) is often computed in log form to select a model from a group of models by picking the model with highest approximate log posterior probability.

$$\log\left(p\left(M_i|y\right)\right) \approx \log\left(\ell_i\left(\hat{\theta}_{MLE,i}\right)\right) - \frac{p_i}{2}\log(n) \tag{6.3.7}$$

---

density.

(6.3.7) is often called the Bayeisan Information Criterion (BIC) or Schwarz criterion. The prior has no influence on the BIC expression. In some situations, the BIC is used to compute an approximation to the Bayes Factor.

$$\begin{aligned} \text{BF} &\approx \log\left(\ell_1\left(\hat{\theta}_1\right)\right) - \log\left(\ell_2\left(\hat{\theta}_2\right)\right) - \left(\frac{p_1 - p_2}{2}\right)\log(n) \\ &= \log\left(\text{LR}_{1,2}\right) - \frac{\Delta p}{2}\log(n) \end{aligned} \tag{6.3.8}$$

The BIC can be extremely inaccurate and should be avoided whenever possible. However, the expression in (6.3.8) is useful to illustrate a fundamental intuition for Bayes Factors. The posterior model probability includes an "automatic" or implicit penalty for models which have higher dimensional parameters. The BF recognizes that adding parameters can simply "overfit" the data and this is automatically accounted for without resort to ad hoc procedures such as out-of-sample validation.

6.4 Computing Bayes Factors Using Importance Sampling

The marginal density of the data for each model can be written as the integral of the un-normalized posterior over the parameter space as in (6.1.8). We can apply importance sampling techniques to this problem (as in Gelfand and Dey, eqn (23)).

$$p\left(y|M_i\right) = \int \tilde{p}\left(\theta|y\right)d\theta = \int \frac{\tilde{p}\left(\theta|y\right)}{q(\theta)} q(\theta)\, d\theta \tag{6.4.1}$$

Here $q(\bullet)$ the importance sampling density. We note that, unlike applications of importance sampling to computing posterior moments, we require the full importance sampling density form, including normalizing constants. Using draws from the importance sampling density, we can estimate the BF as a ratio of integral estimates. We note that a separate importance density will be required for both models.

(6.4.2)
$$\hat{BF} = \frac{\frac{1}{R}\sum_r w_r(M_1)}{\frac{1}{R}\sum_r w_r(M_2)}$$
$$w_r(M_i) = \frac{\tilde{p}\left(\theta_i^r \mid y, M_i\right)}{q_i\left(\theta_i^r\right)};\ \theta_i^r \sim q_i$$

Choice and calibration of the Importance Sampling density is critical to the accuracy of the importance sampling approach. Two suggestions can be helpful. First, we should transform to a parameterization that is unrestricted in order to use an elliptically symmetric importance density. For example, we should transform variance/scale parameters to an unrestricted parameterization. For the scalar case, we can write $\theta = \exp(\gamma)$ and use the Jacobian, $\exp(\gamma)$. For covariance matrices, we can transform to the non-diagonal elements of the Cholesky root and exp of the diagonal elements.

(6.4.3)
$$\Sigma = U'U$$
$$U = \begin{bmatrix} e^{\gamma_{1,1}} & \gamma_{1,2} & \cdots & \gamma_{1,p} \\ 0 & e^{\gamma_{2,2}} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \gamma_{p-1,p} \\ 0 & \cdots & 0 & e^{\gamma_{p,p}} \end{bmatrix}$$

(6.4.4)
$$J(\Gamma) = 2^p \prod_{i=1}^{p} e^{\gamma_{ii}(p-i+1)} \prod_{i=1}^{p} e^{\gamma_{ii}} = 2^p \prod_{i=1}^{p} e^{\gamma_{ii}(p+2-i)}$$

Typically, we would use a Mvst importance density with a moderate degrees of freedom (to insure fatter tails than the un-normalized posterior) and use the MCMC draws from the posterior to assess a mean and covariance matrix in the transformed parameters.

$$\gamma \sim \text{Mvst}\left(\upsilon, \overline{\gamma}, s^2\hat{V}\right)$$

(6.4.5)
$$\overline{\gamma} = \tfrac{1}{R}\textstyle\sum_r \gamma^r_{\text{mcmc}}$$

$$\hat{V} = \tfrac{1}{R}\textstyle\sum_r \left(\gamma^r_{\text{mcmc}} - \overline{\gamma}\right)\left(\gamma^r_{\text{mcmc}} - \overline{\gamma}\right)'$$

The importance density would be tuned by selecting the constant s to insure that the distribution of the importance weights is reasonable (not driven by a small number of outliers). The importance sampling estimate of model probability would now be expressed as

(6.4.6)
$$\hat{p}(y) = \tfrac{1}{R}\textstyle\sum_r \dfrac{\ell\left(\theta\left(\gamma^r\right)\right)p\left(\theta\left(\gamma^r\right)\right)J\left(\theta\left(\gamma^r\right)\right)}{q\left(\gamma^r\right)}$$

6.5 Bayes Factors Using MCMC Draws from the Posterior

Typically, we have an MCMC method implemented for each model under consideration. This gives us the ability to simulate from the posterior distribution of each model's parameters. Therefore, there is a natural interest in methods which can express Bayes Factors as the expectation of quantities with respect to the posterior distribution. These identities allow for the "re-use" of already existing posterior draws for the purpose of estimating the posterior expectation.

Gelfand and Dey (1994) provide one such basic identity.

(6.5.1)
$$\int \frac{q(\theta)}{\tilde{p}\left(\theta|y, M_i\right)} p\left(\theta|y, M_i\right) d\theta = \frac{1}{p\left(y|M_i\right)}$$

(6.5.1) can be verified as follows:

$$\int \frac{q(\theta)}{\tilde{p}\left(\theta|y, M_i\right)} p\left(\theta|y, M_i\right) d\theta = \int \frac{q(\theta)}{p\left(y|M_i\right)p\left(y|\theta, M_i\right)} p\left(\theta|y, M_i\right) d\theta$$

$$= \frac{1}{p\left(y|M_i\right)} \int q(\theta)\, d\theta = \frac{1}{p\left(y|M_i\right)}$$

This derivation makes it clear that q must be a proper density. (6.5.1) can be used to express the Bayes Factor as a ratio of posterior expectations.

$$\text{(6.5.2)} \qquad \mathrm{BF}\left(\mathrm{M}_1 \text{ vs. } \mathrm{M}_2\right) = \frac{\mathrm{p}\left(\mathrm{y}\middle|\mathrm{M}_1\right)}{\mathrm{p}\left(\mathrm{y}\middle|\mathrm{M}_2\right)} = \frac{\mathrm{E}_{\theta|\mathrm{y},\mathrm{M}_2}\left[\dfrac{\mathrm{q}_2(\theta)}{\tilde{\mathrm{p}}\left(\theta\middle|\mathrm{y},\mathrm{M}_2\right)}\right]}{\mathrm{E}_{\theta|\mathrm{y},\mathrm{M}_1}\left[\dfrac{\mathrm{q}_1(\theta)}{\tilde{\mathrm{p}}\left(\theta\middle|\mathrm{y},\mathrm{M}_1\right)}\right]}$$

We can estimate each of the marginal densities of the data by

$$\text{(6.5.3)} \qquad \hat{\mathrm{p}}\left(\mathrm{y}\middle|\mathrm{M}_\mathrm{i}\right) = \frac{1}{\frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \dfrac{\mathrm{q}_\mathrm{i}\left(\theta^\mathrm{r}\right)}{\ell\left(\theta^\mathrm{r}\middle|\mathrm{M}_\mathrm{i}\right)\mathrm{p}\left(\theta^\mathrm{r}\middle|\mathrm{M}_\mathrm{i}\right)}}$$

It should be noted that for some models evaluation of the likelihood can be computationally demanding and (6.5.3) will require many thousands of likelihood evaluations.

The $\mathrm{q}(\bullet)$ function above plays a role analogous to the reciprocal of an importance function. As with an importance function, it is important that the q function "match" or mimic the posterior as closely as possible. This will minimize the variance of the "weights"

$$\hat{\mathrm{p}}\left(\mathrm{y}\middle|\mathrm{M}_\mathrm{i}\right) = \frac{1}{\frac{1}{\mathrm{R}}\sum_{\mathrm{r}=1}^{\mathrm{R}} \mathrm{w}^\mathrm{r}\left(\mathrm{M}_\mathrm{i}\right)}$$

$$\mathrm{w}^\mathrm{r}\left(\mathrm{M}_\mathrm{i}\right) = \frac{\mathrm{q}_\mathrm{i}\left(\theta^\mathrm{r}\right)}{\ell\left(\theta^\mathrm{r}\middle|\mathrm{M}_\mathrm{i}\right)\mathrm{p}\left(\theta^\mathrm{r}\middle|\mathrm{M}_\mathrm{i}\right)}$$

However, the desirable tail behavior for the q function is exactly the opposite of that of an importance density. The tails of the q function serve to attenuate the influence of small values of the posterior density on the estimator. Because of the reciprocal formula, small values of the posterior density can create an estimator with infinite variance. On a practical level, a few "outliers" in the $\left\{\theta^\mathrm{r}\right\}$ can dominate all other draws in the estimate of the marginal data density in (6.5.3). For this reason, it is important to choose a q function with

thin tails relative to the posterior. For problems in high dimensions, it may be difficult to select a q function that works well in the sense of matching the posterior while still having thin tails.

A special case of (6.5.2)/(6.5.3) is the estimator of Newton-Raftery (1994) where $q(\theta) = p(\theta|M_i)$.

(6.5.4) $$\hat{p}(y|M_i) = \frac{1}{\frac{1}{R}\sum_{r=1}^{R} \frac{1}{\ell(\theta^r|M_i)}}$$

(6.5.4) is the harmonic mean of the likelihood values evaluated at the posterior draws. The function, `logMargDenNR`, in ***bayesm*** computes this estimator. Thus, only the likelihood must be evaluated to compute the N-R estimate. Many researchers examine the sequence plot of the log-likelihood over the MCMC draws as an informal check on the model fit and convergence. The N-R estimate uses these same likelihood evaluations. In our experience, the sequence plots of log-likelhood values for two or more competing models can be more informative than the computation of Bayes Factors via the N-R method.

The N-R estimate has been criticized as having undesirable sampling properties. In many applications, only handful of draws determine the value of the N-R estimate. If the data is not very informative about the parameters and vague or relative diffuse priors are used, then some of the posterior draws can give rise to very small values of the likelihood and make the N-R estimate unstable. More carefully assessed informative priors can improve the performance of the N-R and Gelfand-Dey estimates.

Jacquier and Polson (2002) provide another useful identity for the computation of Bayes Factors from MCMC output. They consider the case of nested models.

$\theta' = \left(\theta_1', \theta_2'\right)$

$M_0 : \theta_1 = \theta_1^h$

M_1 : unrestricted

We can write the BF as

$$\text{BF} = \frac{p\left(y|M_0\right)}{p\left(y|M_1\right)} = \frac{\int p\left(y|\theta_2, M_0\right) p\left(\theta_2|M_0\right) d\theta_2}{p\left(y|M_1\right)} \tag{6.5.5}$$

If the BF expression in (6.5.5) is integrated over the marginal prior distribution of θ_1 under M_1, we will still obtain the BF as it is a constant not dependent on model parameters.

$$\text{BF} = \int \left[\frac{\int p\left(y|\theta_2, M_0\right) p\left(\theta_2|M_0\right) d\theta_2}{p\left(y|M_1\right)} \right] p\left(\theta_1|M_1\right) d\theta_1 \tag{6.5.6}$$

Using the relationship, $p\left(y|M_1\right) = \frac{p\left(y|\theta_1, \theta_2, M_1\right) p\left(\theta_1, \theta_2|M_1\right)}{p\left(\theta_1, \theta_2|M_1\right)}$, we can write (6.5.6) as

$$\begin{aligned} \text{BF} = \iint & p\left(y|\theta_2, M_0\right) p\left(\theta_2|M_0\right) p\left(\theta_1|M_1\right) \\ & \times \frac{p\left(\theta_1, \theta_2|y, M_1\right)}{p\left(y|\theta_1, \theta_2, M_1\right) p\left(\theta_1, \theta_2|M_1\right)} d\theta_1 d\theta_2 \end{aligned} \tag{6.5.7}$$

Recognizing that $p\left(\theta_1|M_1\right) \Big/ p\left(\theta_1, \theta_2|M_1\right) = 1 \Big/ p\left(\theta_2|\theta_1, M_1\right)$, (6.5.7) becomes

$$\text{BF} = \iint \frac{p\left(y|\theta_2, M_0\right) p\left(\theta_2|M_0\right)}{p\left(y|\theta_1, \theta_2, M_1\right) p\left(\theta_2|\theta_1, M_1\right)} p\left(\theta_1, \theta_2|y, M_1\right) d\theta_1 d\theta_2 \tag{6.5.8}$$

We note that $p\left(y|\theta_2, M_0\right) = p\left(y|\theta_1, \theta_2, M_1\right)\Big|_{\theta_1 = \theta_1^h}$.

(6.5.8) suggests that we can use the posterior draws to form an estimated BF of the form.

(6.5.9) $$\hat{BF} = \frac{1}{R}\sum_r \frac{p\left(y\middle|\theta_1,\theta_2^r\right)\Big|_{\theta_1=\theta_1^h} p\left(\theta_2^r\middle|M_0\right)}{p\left(y\middle|\theta_1^r,\theta_2^r\right)p\left(\theta_2^r\middle|\theta_1^r,M_1\right)}$$

where $\left\{\theta^r\right\}$ are MCMC draws from the posterior under the unrestricted model. We note that, unlike the Savage-Dickey set-up, the priors under M_0 and M_1 need not be linked via conditioning. What is required, however, is that the conditional prior for $\theta_2|\theta_1$ under M_1 must be available as a normalized density.

6.6 Bridge Sampling Methods

Meng and Wong (1996) provide an identity which links together methods that rely on expectations with respect to the posterior with importance sampling methods. A hybrid procedure which relies on both is termed "bridge sampling." The bridge sampling identity starts with a pair of functions $\alpha(\theta)$ and $q(\theta)$ such that $\int \alpha(\theta)p(\theta|y)q(\theta)d\theta > 0$.

(6.6.1) $$1 = \frac{\int \alpha(\theta)p(\theta|y)q(\theta)d\theta}{\int \alpha(\theta)q(\theta)p(\theta|y)d\theta} = \frac{E_q\left[\alpha(\theta)p(\theta|y)\right]}{E_p\left[\alpha(\theta)q(\theta)\right]}$$

Using the relationship between the marginal density of the data and the un-normalized posterior, we can establish the following identity.

(6.6.2) $$p(y) = \frac{E_q\left[\alpha(\theta)\tilde{p}(\theta|y)\right]}{E_p\left[\alpha(\theta)q(\theta)\right]}$$

We can estimate (6.6.2) by approximating both expectations in the numerator and denominator by iid draws from q and MCMC draws from p.

(6.6.3) $$\hat{p}(y) = \frac{\frac{1}{R_q}\sum \alpha\left(\theta_q^r\right)\tilde{p}\left(\theta_q^r|y\right)}{\frac{1}{R_p}\sum \alpha\left(\theta_p^r\right)q\left(\theta_p^r\right)}$$

Meng and Wong (1996) point out that the Gelfand-Dey estimator and the importance sampling estimator are special cases of bridge sampling with a choice of $\alpha(\bullet)$ to be either the importance density or the un-normalized posterior. They consider the question of an "optimal" (in the sense of smallest MSE in estimation) and provide an iterative scheme for constructing $\alpha(\bullet)$ as an weighted combination of Gelfand-Dey and importance sampling. Fruhwirth-Schnatter (2004) applies this iterative scheme to the construction of BF for mixtures of normals and finds the Meng and Wong estimator to be an improvement over standard procedures.

6.7 Posterior Model Probabilities with Un-identified Parameters

We have considered a number of models with unidentified parameters. We have also seen that, in some cases, it is desirable to navigate in the un-identified parameter space and then margin down or "post-process" the MCMC draws to make inferences regarding the identified parameters. A reasonable question to ask is whether draws of the unidentified parameters can be used to estimate posterior model probabilities. From a purely theoretical point of view, this can be justified. Let us assume that θ is not identified but, $\tau(\theta)$, is.

We can then define a transformation, $g(\theta)$, which partitions the transformed parameters into those which are identified and those which are not.

(6.7.1) $$\delta = g(\theta) = \begin{bmatrix} \upsilon(\theta) \\ \tau(\theta) \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}$$

δ_1 is a subvector of length $k - k_1$ containing the unidentified parameters. We start with a prior over the full vector of parameters, $p_\theta(\bullet)$, and then compute the induced prior over the identified parameters.

(6.7.2)
$$p_{\delta_2}(\delta_2) = \int p_{\delta}(\delta_1, \delta_2)\, d\delta_1$$
$$p_{\delta}(\delta_1, \delta_2) = p_{\theta}(g(\theta))\, J_{\theta \to \delta}$$

We can compute the marginal density of the data in two ways. We can compute the density in the space of unidentified parameters or directly on the identified parameters using the induced prior in (6.7.2).

Working in the identified parameter space,

(6.7.3)
$$p'(y) = \int p(y|\delta_2)\, p_{\delta_2}(\delta_2)\, d\delta_2$$

Now consider the marginal density of the data computed in the full, un-identified parameter space.

$$\begin{aligned} p(y) &= \iint p(y|\delta_1, \delta_2)\, p_{\delta}(\delta_1, \delta_2)\, d\delta_1 d\delta_2 \\ &= \int \left[\int p(y|\delta_2)\, p_{\delta_2}(\delta_2)\, d\delta_2 \right] p(\delta_1|\delta_2)\, d\delta_1 \\ &= \int p'(y)\, p(\delta_1|\delta_2)\, d\delta_1 = p'(y) \end{aligned}$$

Here we are using the fact that, $p(y|\delta_1, \delta_2) = p(y|\delta_2)$ since δ_1 represents the unidentified parameters.

Thus, we are theoretically justified in using MCMC draws from the unidentified parameter space in the methods considered in section 6.6. However, in practice we recognize that the draws of the un-identified parameters may exhibit a great deal of variation, especially in situations with vague or very diffuse priors. Again, there is a pay-off to assessing realistic priors.

6.8 Chib's Method

Chib (1995) proposes a method which uses MCMC output to estimate the marginal density of the data. This method is particularly appropriate for models that have a conjugate structure conditional on the value of augmented latent variables. Chib starts with the basic

identity relating the normalized and un-normalized posteriors. This identity holds for any value of θ, indicated by θ^* below

$$p(y) = \frac{\tilde{p}(\theta^*|y)}{p(\theta^*|y)} = \frac{p(y|\theta^*)p(\theta^*)}{p(\theta^*|y)} \tag{6.8.1}$$

The key insight of Chib (1995) is that, for certain conditionally conjugate models, the denominator of (6.8.1) can be expressed as an average of densities which are known up to and including normalizing constants.

For example, consider the arch-typical data augmentation model.

$$\begin{array}{l} y \mid z \\ z \mid \theta \\ \theta \end{array} \tag{6.8.2}$$

For the model in (6.8.2), we can write the ordinate of the posterior at θ^* as the average of the posterior conditional on the latent z over the marginal posterior distribution of z.

$$\begin{aligned} p(\theta|y) &= \int \frac{p(y,z,\theta)}{p(y)} dz = \int \frac{p(\theta|y,z)p(y,z)}{p(y)} dz \\ &= \int \frac{p(\theta|y,z)p(z|y)p(y)}{p(y)} dz = \int p(\theta|y,z)p(z|y) dz \end{aligned} \tag{6.8.3}$$

(6.8.3) suggests that we can estimate the marginal density of the data as follows:

$$\hat{p}(y) = \frac{p(y|\theta^*)p(\theta^*)}{\hat{p}(\theta^*|y)} = \frac{p(y|\theta^*)p(\theta^*)}{\frac{1}{R}\sum_r p(\theta^*|y,z^r)} \tag{6.8.4}$$

$\{z^r\}$ are draws from the marginal posterior of the latent variables. θ^* is usually taken to be the posterior mean or mode, computed from the MCMC draws. (6.8.4) requires that we simply save the latent draws from our MCMC run and that we be able to evaluate the data

density (likelihood) and the prior densities with all normalizing constants. We should note that (6.8.4) requires evaluation of the marginal likelihood (without the latents). For some models, this can be computationally challenging. However, we should point out that the Chib method only requires one likelihood evaluation whereas Gelfand/Dey style methods would require R evaluations.

The Chib method in (6.8.4) requires that we be able to evaluate, $p\left(\theta^{*}|y,z\right)$, including all normalizing constants. For some models, we can use the fact that $p\left(\theta|y,z\right)=p\left(\theta|z\right)$ and that we have a conjugate set-up conditional on z. However, in other applications, this will not be possible. In some cases, $\theta|z$ is not fully conjugate but can be broken into two conjugate blocks. Consider the case where θ is partitioned into $\left(\theta_1,\theta_2\right)$. We can then estimate $p\left(\theta^{*}|y,z\right)$ using the identity, $p\left(\theta_1,\theta_2|y\right)=p\left(\theta_1|y\right)p\left(\theta_2|\theta_1,y\right)$. We can compute the marginal posterior density of θ_1 by averaging the conditional density.

(6.8.5) $$p\left(\theta_1|y\right)=\int p\left(\theta_1|\theta_2,z,y\right)p\left(\theta_2,z|y\right)dz\,d\theta_2$$

(6.8.5) can be estimated by averaging the conditional density over the MCMC draws of θ_2 and z.

(6.8.6) $$\hat{p}\left(\theta_1^{*}|y\right)=\frac{1}{R}\sum_r p\left(\theta_1^{*}\middle|\theta_2^{r},z^{r}\right)$$

However, the conditional posterior density of $\theta_2|\theta_1$ cannot be estimated by averaging the conditional with respect to the marginal posterior as in (6.8.5).

(6.8.7) $$p\left(\theta_2|\theta_1,y\right)=\int p\left(\theta_2|\theta_1,z,y\right)p\left(z|\theta_1,y\right)dz$$

To estimate this density at the point $\left(\theta_1^*, \theta_2^*\right)$ requires a modified MCMC sampler for $\left(z, \theta_2\right)$ given $\theta_1 = \theta_1^*$. As Chib points out, this is simple to achieve by shutting down the draws for θ_1. If $z_{\theta_1^*}^r$ are draws from the marginal posterior of z given $\theta_1 = \theta_1^*$, then we can estimate (6.8.7) by

$$\hat{p}\left(\theta_2^* \middle| \theta_1^*, y\right) = \frac{1}{R}\sum_r p\left(\theta_2^* \middle| \theta_1^*, z_{\theta_1^*}^r, y\right) \tag{6.8.8}$$

To estimate the marginal density of the data, we put together (6.8.6) and (6.8.8).

$$\begin{aligned} \hat{p}(y) &= \frac{p\left(y \middle| \theta^*\right) p\left(\theta^*\right)}{\hat{p}\left(\theta_1^* \middle| y\right) \hat{p}\left(\theta_2^* \middle| \theta_1^*, y\right)} \\ &= \frac{p\left(y \middle| \theta^*\right) p\left(\theta^*\right)}{\left(\frac{1}{R}\sum_r p\left(\theta_1^* \middle| \theta_2^r, z^r\right)\right)\left(\frac{1}{R}\sum_r p\left(\theta_2^* \middle| \theta_1^*, z_{\theta_1^*}^r, y\right)\right)} \end{aligned} \tag{6.8.9}$$

For cases in which there is no convenient decomposition into conjugate blocks, the method of Chib and Jeliazhov (2001) can be used.

6.9 An Example of Bayes Factor Computation: Diagonal MNP models

We will illustrate the methods of importance sampling, Newton-Raftery, and Chib using an example of comparison of different MNP models. The estimation of the off-diagonal covariance (correlation) elements in the MNP is often difficult. We typically find that the diagonal elements (relative variances) are much more precisely estimated. In addition, we often find large differences in relative variance between choice alternatives. This is consistent with a view that X variables explain different portions of the utility of each choice. For example, some choice alternatives may have utility explained well by price while others may have utility that depends on attributes not measured in our data. For this reason, the

MNP model with a non-scalar but diagonal covariance matrix could be considered a central model. We will develop Bayes Factors to compare the diagonal MNP with an "identity" MNP or a model with an Identity covariance matrix. This would closely approximate the IIA properties of the MNL model.

$$y_i = \sum_{j=1}^{p} I\left(z_{i,j} = \max\left(z_i\right)\right)$$
$$z_i = X_i\beta + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, \Lambda\right)$$

Diagonal MNP:
$$\Lambda = \begin{bmatrix} 1 & & & \\ & \sigma_{22} & & \\ & & \ddots & \\ & & & \sigma_{pp} \end{bmatrix}$$

$$y_i = \sum_{j=1}^{p} I\left(z_{i,j} = \max\left(z_i\right)\right)$$

Identity MNP:
$$z_i = X_i\beta + \varepsilon_i$$
$$\varepsilon_i \sim N\left(0, I_p\right)$$

We use normal priors on β and independent scaled inverted chi-squared priors on the diagonal elements 2, …, p. The first diagonal element is fixed at 1.

(6.9.1)
$$\beta \sim N\left(\bar{\beta}, A^{-1}\right)$$
$$\sigma_{j,j} \sim_{\text{ind}} \upsilon_0 s_0^2 / \chi^2_{\upsilon_0}$$

To apply the Newton-Raftery method, we must implement MCMC samplers for each of the two MNP models and compute the likelihood at each draw as in (6.5.4). The diagonal MNP sampler and Identity MNP samplers are special cases of the algorithms given in chapter 4, section 2. To compute the likelihood, we must compute the choice probabilities for each observation. The general MNP likelihood requires evaluation of the integral of a correlated normal random variable over a rectangular region as discussed in section 5.2.1. However, for the diagonal MNP model, the choice probabilities can be

simplified to the average of normal cdfs with respect to a univariate normal distribution. Both importance sampling and N-R methods will be computationally intensive due to the evaluation of the likelihood over the set of MCMC draws or over draws from the importance density. This argues in favor of the Chib method which only requires one evaluation of the likelihood.

To implement an importance function approach, we first transform the variance parameters to an unrestricted space. This only applies to the Diagonal MNP model. If θ is the stacked vector of the parameters of the Diagonal MNP model, $\theta' = \left(\beta, \text{diag}\left(\Lambda\right)'\right)$, then we define the transformed vector η by

$$\eta = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} \beta \\ \ln\left(\text{diag}\left(\Lambda\right)\right) \end{bmatrix}; \ \theta = \begin{bmatrix} \eta_1 \\ e^{\eta_2} \end{bmatrix} \tag{6.9.2}$$

with Jacobian:

$$J_{\theta \to \eta} = \left\| \begin{matrix} I_{\dim(\beta)} & & 0 \\ & e^{\eta_{2,2}} & \\ 0 & & \ddots \\ & & & e^{\eta_{2,p}} \end{matrix} \right\| = \prod_{i=2}^{p} e^{\eta_{2,i}} \tag{6.9.3}$$

To implement the importance sampling method, we use a normal importance function with location and scale chosen using the MCMC draws in the transformed parameter η.

$$\begin{aligned} q\left(\eta\right) &= \phi\left(\bar{\eta}, s^2 \hat{V}\right) \\ \bar{\eta} &= \tfrac{1}{R} \textstyle\sum_r \eta^r_{\text{mcmc}} \\ \hat{V} &= \tfrac{1}{R} \textstyle\sum_r \left(\eta^r_{\text{mcmc}} - \bar{\eta}\right)\left(\eta^r_{\text{mcmc}} - \bar{\eta}\right)' \end{aligned} \tag{6.9.4}$$

where ϕ denotes a normal density. We approximate the marginal density of the data using draws from q as follows:

$$\hat{p}(y) = \frac{1}{N}\sum_i \frac{p\left(y\middle|\theta\left(\eta^i\right)\right)p\left(\theta\left(\eta^i\right)\right)J_{\theta\to\eta}\left(\theta\left(\eta^i\right)\right)}{q\left(\eta^i\right)} = \frac{1}{N}\sum_i w_i \tag{6.9.5}$$

We note that the q density in the denominator of (6.9.5) is the full normalized density. The importance sampling method can be tuned by choice of the scaling parameter, s.

The Chib method can be implemented directly using (6.8.4) for the Identity MNP model. However, for the Diagonal MNP model, we use the variant of the Chib method which breaks the parameter vector into two parts, $\theta_1 = \beta$, $\theta_2 = \mathrm{diag}(\Lambda)$, in equation (6.8.9).

As an illustration, we simulate data from a Diagonal MNP model with p = 3 and $\mathrm{diag}(\Lambda) = (1,2,3)$, N=500. X contains p-1 intercepts and two regressors which are simulated unif(-1,1). We compute the BF factors for the Diagonal MNP vs. the Identity MNP. We use 150,000 draws from the MCMC samplers and the importance sampling density. We assess modestly informative priors with $\bar{\beta} = 0$, A=.25I, $\upsilon_0 = 10$, and $s_0^2 = 1$.

Figure 6.1 shows the log-likelihood values for the Diagonal MNP (dark) and the Identity MNP (light) plotted for every 75th draw. The figure clearly shows that the Diagonal MNP fits the data better than the Identity MNP. Even though inspection of this plot is somewhat "informal," we highly recommend this as a rough assessment of the models as well as the convergence of the MCMC algorithm. In fact, in our experience this sort of plot can be more informative than formal Bayes Factor computations.

Below we show the results of Bayes Factor computations using all three methods.

Method	BF
NR	3.4
IS (s=1.5)	2.9
IS (s=2.0)	2.9
Chib	3.4

These numbers do not convey the sampling error in each of the estimates. As an informal assessment, we made more than 10 different runs and observed very little variation in the IS sampling or Chib numbers and a wide range of NR BFs from 2.0 to 3.5. We note that our prior is more informative than most and we expect that with more traditional "vague" prior settings we would see even more variation in the NR numbers.

The NR approach is frequently criticized in the literature due to the fact that the estimator has an infinite variance under some conditions. However, the convenience of the NR approach accounts for its widespread popularity. Figure 6.2 illustrates the problems with the NR approach more dramatically. The NR estimate is driven by a small portion of the values of the reciprocal of the likelihood as shown by the histogram of the reciprocal of the likelihood for every 10^{th} draw (right hand side panel). While the distribution of log-likelihood values is not particularly abnormal, the harmonic mean estimator is particularly vulnerable to outlying observations.

The Importance Sampling approach does not have as severe problem with outlying observations as the NR approach. However, even with careful choice of scaling constant[3], the weights have outliers as shown in figure 6.3 Figure 6.3 shows the distribution of the importance sampling weights normalized by the median of the weight. This figure does not instill confidence in the importance sampling approach either. The parameter space is only

[3] In this application, choice of the scaling constant is critical. We experimented with student t importance densities and found little benefit.

of dimension seven in this example and we might expect the problems with the importance sampling approach to magnify in larger dimension parameter spaces.

The Chib approach to calculating BF relies on various estimates of the ordinate of posterior or conditional posterior densities. For the Diagonal MNP, the Chib approach averages $p\left(\beta^* \middle| \Lambda, z, y\right)$ over MCMC draws of Λ and z. $p\left(\Lambda^* \middle| \beta^*, z, y\right)$ is averaged over draws of $z\middle|\beta^*, y$. For the Identity MNP, only the density of β is averaged. Figure 6.4 shows the distribution of these densities (right hand side) and the log of these densities. Again, there are very large outliers which influence the averages. These outliers contribute to instability of this approach.

In this example, the Importance Sampling approach provides the most reliable results. However, it is important to note that all three approaches are sensitive to outliers. The "numerical standard error" formulas used to measure the sampling errors in the IS and Chib approach are unlikely to be reliable in the presence of such large outliers.

6.10 Marketing Decisions and Bayesian Decision Theory

Computation of posterior model probabilities is motivated as a special case of Bayesian decision theory with a zero-one loss function. In marketing problems, the profit function of the firm provides a more natural choice of loss. In addition, there is often considerable parameter or modeling uncertainty. Bayesian decision theory is ideally suited for application to many marketing problems in which a decision must be made given substantial parameter or modeling uncertainty. In these situations, the uncertainty must factor into the decision itself. The marketing decision maker takes an action by setting the value of various variables designed to quantify the marketing environment facing the consumer (such as price or advertising levels). These decisions should be affected by the level of uncertainty facing the

marketer. To make this concrete, begin with a probability model that specifies how the outcome variable (y) is driven by the explanatory variables (x) and parameters θ.

$$p(y|x,\theta)$$

The decision maker has control over a subset of the x vector, $x' = \left[x_d', x_{cov}' \right]$. x_d represents the variables under the decision maker's control and x_{cov} are the covariates. The decision maker chooses x_d so as to maximize the expected value of profits where the expectation is take over the distribution of the outcome variable. In a fully Bayesian Decision Theoretic treatment, this expectation is taken with respect to the posterior distribution of θ as well as the predictive conditional distribution $p(y|x_d, x_{cov})$.

$$\begin{aligned} \pi^*(x_d|x_{cov}) &= E_\theta\left[E_{y|\theta}\left[\pi(y|x_d) \right] \right] \\ &= E_\theta\left[\int \pi(y|x_d) p(y|x_d, x_{cov}, \theta) dy \right] \\ &= E_\theta\left[\bar{\pi}(x_d|x_{cov}, \theta) \right] \end{aligned} \tag{6.10.1}$$

The decision maker chooses x_d to maximize profits π^*. In general, the decision maker can be viewed as minimizing expected loss which is frequently taken as – profits but need not be in all cases (see, for example, Steenburgh et al (2002))

Plug-in Vs. Full Bayes Approaches

The use of the posterior distribution of the model parameters to compute expected profits is an important aspect of the Bayesian approach. In an approximate or conditional Bayes approach, the integration of the profit function with respect to the posterior distribution of θ is replaced by an evaluation of the function at the posterior mean or mode of the parameters. This approximate approach is often called the "plug-in" approach, or according to Morris (1983), "Bayes Empirical Bayes".

(6.10.2) $$\pi^{*}\left(x_d\right) = E_{\theta|y}\left[\overline{\pi}\left(x_d\left|\theta\right.\right)\right] \neq \overline{\pi}\left(x_d\left|\hat{\theta} = E_{\theta|y}\left[\theta\right]\right.\right)$$

When the uncertainty in θ is large and the profit function is non-linear, errors from the use of the plug-in method can be large. In general, failure to account for parameter uncertainty will overstate the potential profit opportunities and lead to "over-confidence" that results in an overstatement of the value of information (c. f. Montgomery and Bradlow 1999).

Use of Alternative Information Sets

One of the most appealing aspects of the Bayesian approach is the ability to incorporate a variety of different sources of information. All adaptive shrinkage methods utilize the similarity between cross-sectional units to improve inference at the unit level. A high level of similarity among units leads to a high level of information shared. Since the level of similarity is determined by the data via the first-stage prior, the shrinkage aspects of the Bayesian approach adapt to the data. For example, Neelameghan and Chintagunta (1999) show that similarities between countries can be used to predict the sales patterns following the introduction of new products.

The value of a given information set can be assessed using a profit metric and the posteriors of θ corresponding to the two information sets. For example, consider two information sets A and B along with corresponding posteriors, $p_A(\theta), p_B(\theta)$. We solve the decision problem using these two posterior distributions.

(6.10.3) $$\Pi_l = \max_{x_d} \pi_l^{*}\left(x_d\left|x_{cov}\right.\right) = \max_{x_d} \int \overline{\pi}\left(x_d\left|x_{cov},\theta\right.\right) p_l(\theta)\, d\theta \qquad l = A, B$$

We now turn to the problem of valuing disaggregate information.

Valuation of Disaggregate Information

Once a fully decision-theoretic approach has been specified, we can use the profit metric to value the information in disaggregate data. We compare profits that can be obtained via our disaggregate inferences about $\{\theta_i\}$ with profits that could be obtained using only aggregate information. The profit opportunities afforded by disaggregate data will depend both on the amount of heterogeneity across the units in the panel data as well as the level of information at the disaggregate level.

To make these notions explicit, we will lay out the disaggregate and aggregate decision problems. As emphasized in chapter 5, Bayesian methods are ideally suited for inference about the individual or disaggregate parameters as well as the common parameters. Recall the profit function for the disaggregate decision problem.

$$\pi_i^*\left(x_{d,i}\middle|x_{cov,i}\right)=\int\bar{\pi}\left(x_{d,i}\middle|x_{cov,i},\theta_i\right)p\left(\theta_i\middle|\text{Data}\right)d\theta_i \tag{6.10.4}$$

Here we take the expectation with respect to the posterior distribution of the parameters for unit "i". Total profits from the disaggregate data are simply the sum of the maximized values of the profit function above. $\Pi_{disagg}=\sum\pi_i^*\left(\tilde{x}_{d,i}\middle|x_{cov,i}\right)$ where $\tilde{x}_{d,i}$ is the optimal choice of $x_{d,i}$

Aggregate profits can be computed by maximizing the expectation of the sum of the disaggregate profit functions with respect to the predictive distribution of θ_i

$$\begin{aligned}&\pi_{agg}\left(x_d\right)=E_\theta\left[\sum\bar{\pi}\left(x_d\middle|x_{cov,i},\theta\right)\right]=\int\sum\bar{\pi}\left(x_d\middle|x_{cov,i},\theta\right)\bar{p}\left(\theta\right)d\theta\\&\Pi_{agg}=\pi_{agg}(\tilde{x}_d)\end{aligned} \tag{6.10.5}$$

The appropriate predictive distribution of θ, $\bar{p}(\theta)$, is formed from the marginal of the first stage prior with respect to the posterior distribution of the model parameters.

$$\bar{p}(\theta) = \int p(\theta|\tau)p(\tau|\text{Data})d\tau$$

Comparison of Π_{agg} with Π_{disagg} provides a metric for the achievable value of the disaggregate information.

6.11 An Example of Bayesian Decision Theory: Valuing Household Purchase Information

As emphasized in section 6.10, valuation of information must be made within a decision context. Rossi et al (1996) consider the problem of valuing household purchase information using a targeted couponing problem. In traditional couponing exercises, neither retailers nor manufacturers have access to information about individual consumers. Instead, a "blanket" coupon is distributed (via mass mailings or inserts in newspapers). This blanket coupon is, in principle, available to all consumers (note: the large denomination "rebates" available on consumer durable goods are the same idea with a different label). In order to be profitable, the issuer of the coupon relies on a indirect method of price discrimination where the consumers with lower willingness to pay use the coupon with higher probability that those with higher willingness to pay. In the late 1980s, the technology for issuing customized coupons became available. For example, Catalina Marketing Incorporated started a highly successful business by installing coupon printers in grocery stores that were connected to the point of sale terminal. This technology opened the possibility of issuing coupons based on purchase history information. At the most elementary level, it was now possible to issue coupons to consumers who exhibited interest in a product category by purchasing some product in this category. For example, an ice cream manufacturer could pay Catalina to issue coupons for its ice cream products to anyone observed to be purchasing ice cream.

Frequent shopper programs adopted by many retailers also provide a source of purchase and demographic information. By linking purchase records via the frequent shopper id, it is possible to assemble a panel of information about a large fraction of the retailers' customers. This allows for much more elaborate coupon trigger strategies which involve estimated willingness to pay. That is, we might be able to directly estimate a consumers' willingness to pay for a product by estimating a demand model using purchase history data. In addition, frequent shopper programs sometimes collect limited demographic information either directly on enrollment applications or via indirect inference from the members address (so-called geo-demographic information).

These developments mean that the issuer of the coupon now has access to a rich information set about individual consumers. How much should the issuer be willing to pay for this information? Or, what is the value of targeted couponing relative to the traditional blanket approach? To answer these questions, a model of purchase behavior is required as well as a loss function. The natural loss function for this problem is expected incremental profits from the issue of a coupon. Rossi et al postulate a hierarchical probit model similar to the hierarchical logit model of chapter 5 in which demographic information enters the hierarchy. The hierarchical probit model allows the issue of the coupon to make inferences about consumer level purchase probabilities given various information sets. Clearly, only a very limited set of information is available for each consumer. This means that these inferences will imprecise and the degree of "targeting" or customization of the face value of the coupon will depend not only on how different consumers are in willingness to pay but also on how precisely we can measure these differences. This full accounting for parameter uncertainty is a key feature of a Bayesian decision-theoretic approach. This will guard

against the over-confidence that can arise from "plug-in" approaches that simply estimate parameters without accounting for uncertainty.

Rossi et al start with a multinomial probit model at the household or consumer level. Households are confronted with a choice from among p brands with covariates taken to be measures of the marketing mix facing the consumer (price and advertising variables).

(6.11.1) $$y_{h,t} = j \text{ if } \max\left(z_{h,t} = X_{h,t}\beta_h + \varepsilon_{h,t}\right) = z_{j,h,t}$$

(6.11.2) $$\varepsilon_{h,t} \sim N(0, \Lambda)$$

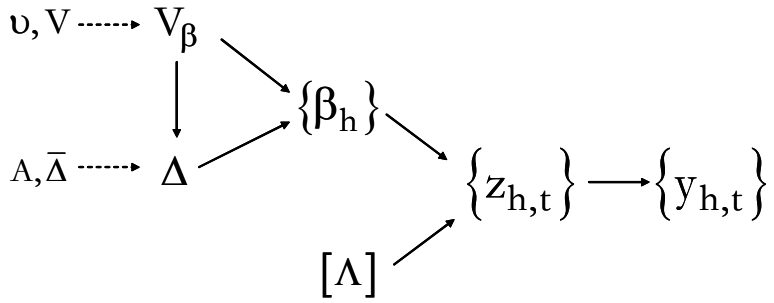
In Rossi et al (in contrast to the standard MNP in 4.2), the covariance matrix of the latent errors (Λ) is taken to be diagonal. This means that we do not have to difference the system and that identification is achieved by setting $\lambda_{11} = 1$. Rossi et al make this simplification for practical and data-based reasons. Given β_h, Λ, choice probabilities for the diagonal MNP model are simple to compute, requiring only univariate integration. Since these choice probabilities figure in the loss/profit function and will be evaluated for many 1000s of draws, it is important that these probabilities can be computed at low cost. This is not as much of a consideration now as it was in the early 90s when this research was conducted. However, this is not the only reason for using a diagonal covariance. Much of the observed correlation in non-hierarchical probit models can be ascribed to heterogeneity. Once heterogeneity is taken into account via the hierarchy, the errors in the unit-level probit are much less correlated. However, they can often be very heteroskedastic as pointed out in section 6.9.

This "unit-level" model is coupled with the, by now, standard normal model of heterogeneity.

(6.11.3)
$$\beta_h = \Delta' z_h + v_h$$
$$v_h \sim N\left(0, V_\beta\right)$$

z_h is a vector of demographic information. The key task is to compute the predictive distribution of the household parameters and choice probabilities for various information sets. That is we must compute, $p\left(\beta_h, \Lambda \middle| \Omega^*\right)$, where Ω^* denotes a particular information set. This predictive distribution will be used in the decision problem to determine what is the optimal face value coupon to issue to household h.

Given the proper choice of priors on the common parameters, Λ, Δ, V_β, we can define an MCMC method to draw from the joint posterior of $\{\beta_1, \ldots, \beta_h, \ldots, \beta_H\}, \Lambda, \Delta, V_\beta$. Rossi et al observed that one can develop a Gibbs Sampler for this problem by using the McCulloch and Rossi (1994) sampler coupled with the standard normal hierarchical set-up. That is, given $\{\beta_h\}$, we have standard normal, IW draws for Δ, V_β as in section 3.7 or 5.3. A DAG for the model is given below:



Thus, we can obtain draws from the marginal posterior distributions of all household and common parameters.

Consider three information sets:

1. "Full" : purchase history information on household h and demographics

2. "Demographics Only": just knowledge of the demographic vector for household h
3. "Blanket": no information about household h but only information about the "population" distribution of households obtained from a sample of households.

Rossi et al consider a "choices-only" information set which we will not discuss here. We must compute the "predictive" or posterior distribution of β_h for each of the three information sets. The first and richest information set is a natural byproduct of the Gibbs Sampler defined above. That is, when we compute the posterior distribution of each β_h, we will take into account each households demographics and purchase history. In addition, information from other households will influence our views about a particular β_h via inferences about the common parameters. We note that by looking at the marginal distribution of the posterior draws, we are integrating out or averaging over draws of the common parameters, accounting for uncertainty in these quantities. That is,

$$p\left(\beta_h \left| \{y_1,\ldots,y_h,\ldots,y_H,X_1,\ldots,X_h,\ldots,X_H\},Z\right.\right) =$$
$$\int p\left(\beta_h \left| y_h,X_h,\Lambda,\Delta,V_\beta\right.\right) p\left(\Lambda,\Delta,V_\beta \left| \{y_1,\ldots,y_h,\ldots,y_H,X_1,\ldots,X_h,\ldots,X_H\},Z\right.\right) d\Lambda d\Delta dV_\beta$$

Z is the matrix of all household demographics. We cannot simply "plug-in" estimates of the common parameters.

$$p\left(\beta_h \left| \{y_1,\ldots,y_h,\ldots,y_H,X_1,\ldots,X_h,\ldots,X_H\},Z\right.\right) \neq p\left(\beta_h \left| y_h,X_h,\hat{\Lambda},\hat{\Delta},\hat{V}_\beta\right.\right)$$

The second and third information sets require some thought. If we only have demographic information but no purchase information on household h, we then must compute an appropriate predictive distribution from the model in (6.11.3).

$$\int p\left(\beta_h \left| z_h,\Delta,V_\beta\right.\right) p\left(\Delta,V_\beta \left| \text{Info}\right.\right) d\Delta dV_\beta \tag{6.11.4}$$

In order to undertake the computation in (6.11.4), we must specify an information set on which we base our inferences on the parameters of the heterogeneity distribution. Our idea

is that we might have a "pet" panel of households, not including household h, on which we observed purchases and demographics so that we can gauge the distribution of probit parameters. That is, we have a sample of data which enables us to gauge the extent of differences among households in this population. This distribution could also simply reflect prior beliefs on the part of managers regarding the distribution. To implement this, we use the posterior distribution from our sample of households. We can define a simulator for (6.11.4) by using all R draws of Δ, V_β and drawing from the appropriate normal draw.

(6.11.5) $$\beta_h^r \left| \Delta^r, V_\beta^r \sim N\left(\Delta^{r'} z_h, V_\beta^r\right)\right.$$

The third and coarsest information set is the same information set used in setting a blanket coupon face value. No information is available about household h. The only information available is information regarding the distribution of the β parameters and demographics in the population. In this situation, we simply integrate over the distribution of demographics as well as the posterior distribution of Δ, V_β.

(6.11.6) $$\int p\left(\beta_h \left| z_h, \Delta, V_\beta\right.\right) p\left(z_h\right) p\left(\Delta, V_\beta \left| \text{Info}\right.\right) dz_h d\Delta dV_\beta$$

Rossi et al use the empirical distribution of z to perform this integral.

Thus, we now have the ability to compute the predictive distribution of each household parameter vector for each of the three information sets. We must now pose the decision problem. The problem is to choose the face value of a coupon to as to maximize total profits. We model the effect of a coupon as simply a reduction in price for the brand for which the coupon is issued. If the coupon is issued for brand i, the decision problem as be written as follows:

(6.11.7) $$\max_F \pi(F) = \int \Pr\left[i \left| \beta_h, \Lambda, X(F)\right.\right] (M - F) p\left(\beta_h, \Lambda \left| \Omega^*\right.\right) d\beta_h d\Lambda$$

$p\left(\beta_h, \Lambda \middle| \Omega^*\right)$ is the predictive or posterior distribution for information set Ω^*. M is the margin on the sale of brand i without consideration of the coupon cost. X(F) denotes the value of the marketing mix variables with a coupon of face value, F. We assume that the effect of the coupon is to reduce the price of alternative i by F. In many cases, the amount of information available about β_h is very small so that the predictive distribution will be very diffuse. It is therefore, extremely important that the integration in (6.11.7) be performed. If "plug-in" estimates of β_h are used, profits can be dramatically overstated due to overconfidence. At the plug-in estimate, the probabilities can be more extreme, suggesting that the coupon will have a greater effect on purchase behavior. In the case of the most popular Catalina product, coupons are issued on the basis of only *one* purchase observation so the extent of the effect can be huge.

Figure 6.5 shows the distribution of expected revenue for a specific household and various coupons face values in an example drawn from Rossi et al. The product is canned tunafish and the coupon face values are given in cents and restricted to be multiples of 5 cents. The boxplots show the considerable uncertainty regarding expected revenue based on very imprecise inferences about β_h. The solid dots in the figure correspond to predicted revenue as the "plug-in" estimate of β_h equal to the posterior mean. The figure shows the "over-confidence" aspect of plug estimates.

Rossi et al demonstrate that, relative to no household-specific information, various information sets regarding households have the potential for large value. That is, even with a small amount of information, the ability to customize the face value of the coupon is high. Revenues from even one observation are over 50 per cent higher than in the blanket coupon

condition. Revenues from longer purchase histories can be even greater, exceeding 100 per cent larger.

Figure 6.1

Log-likelihood Values: Diagonal Vs. Identity MNP

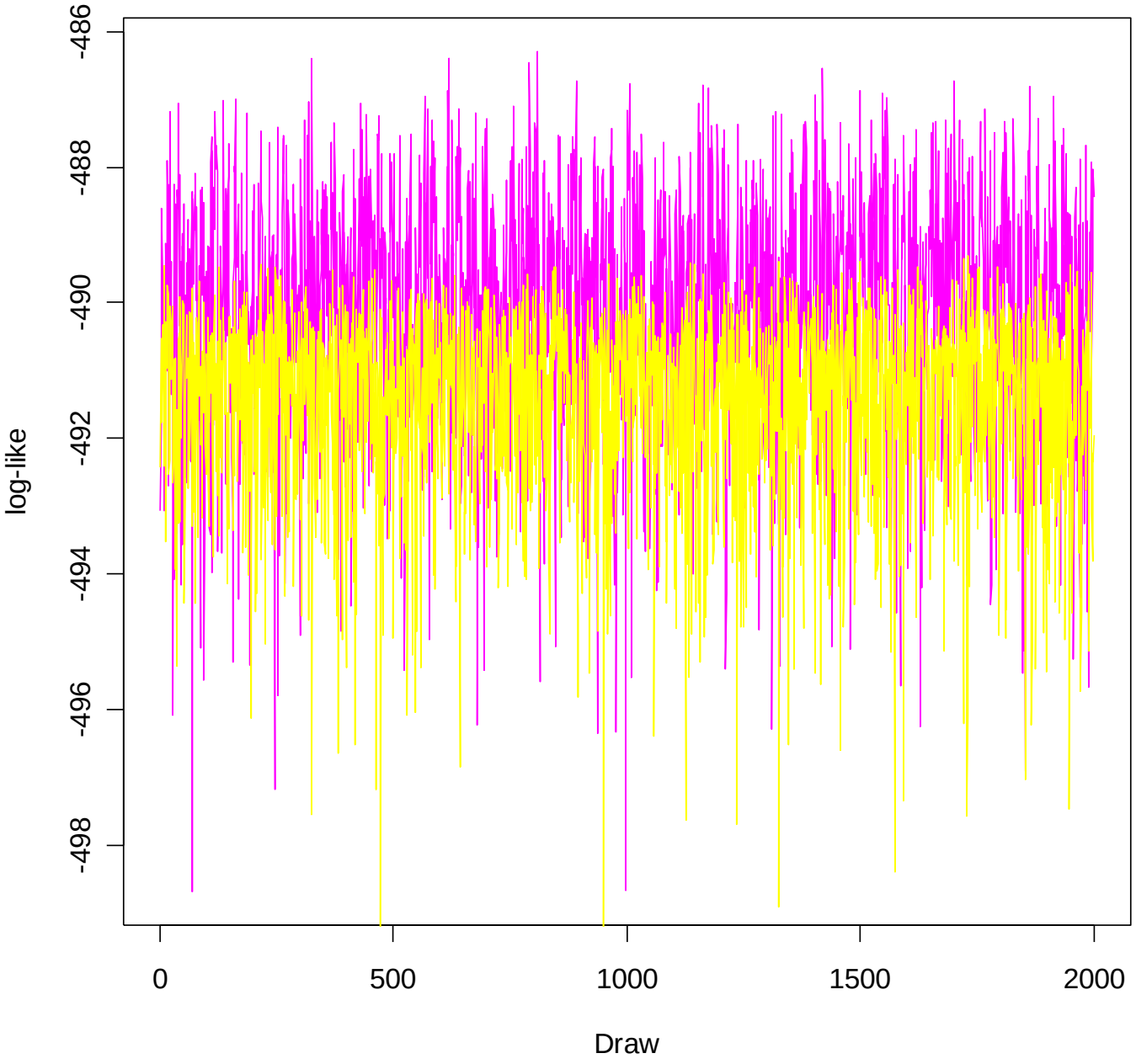


Figure 6.2
Difficulties with Newton-Raftery BF Estimator

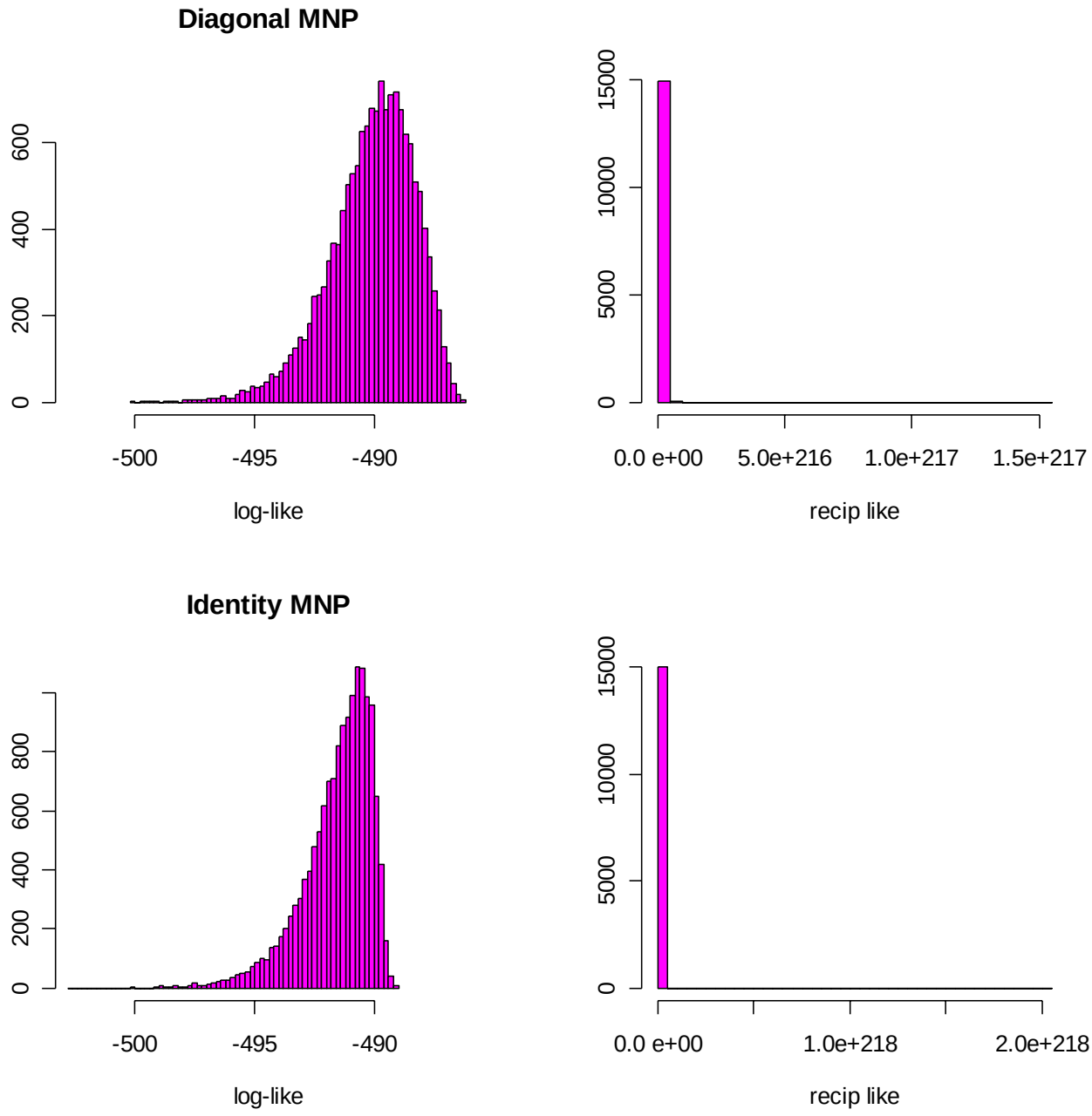


Figure 6.3
Importance Sampling Weight Distribution

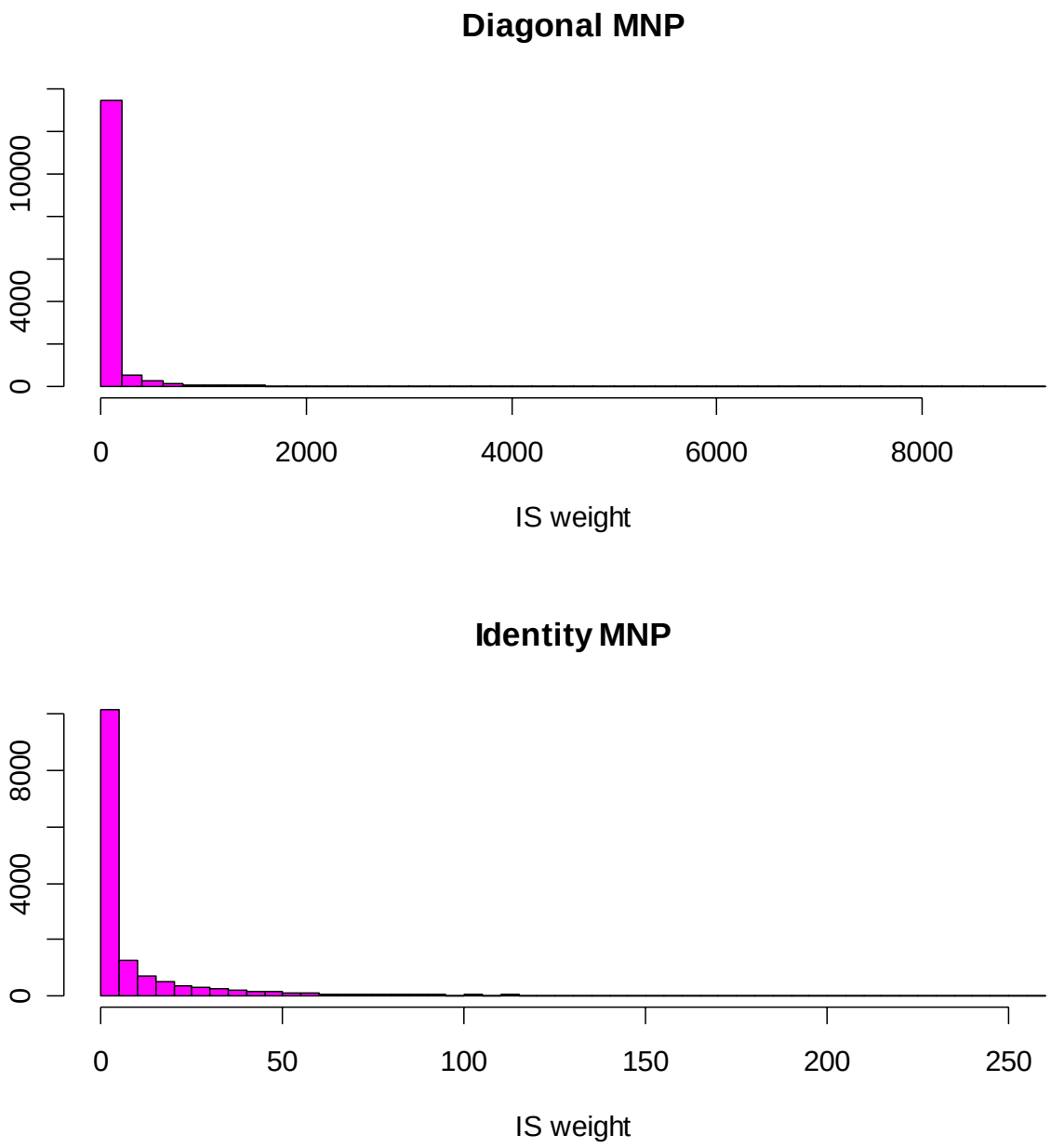


Figure 6.4
Values of Densities in the Chib Approach

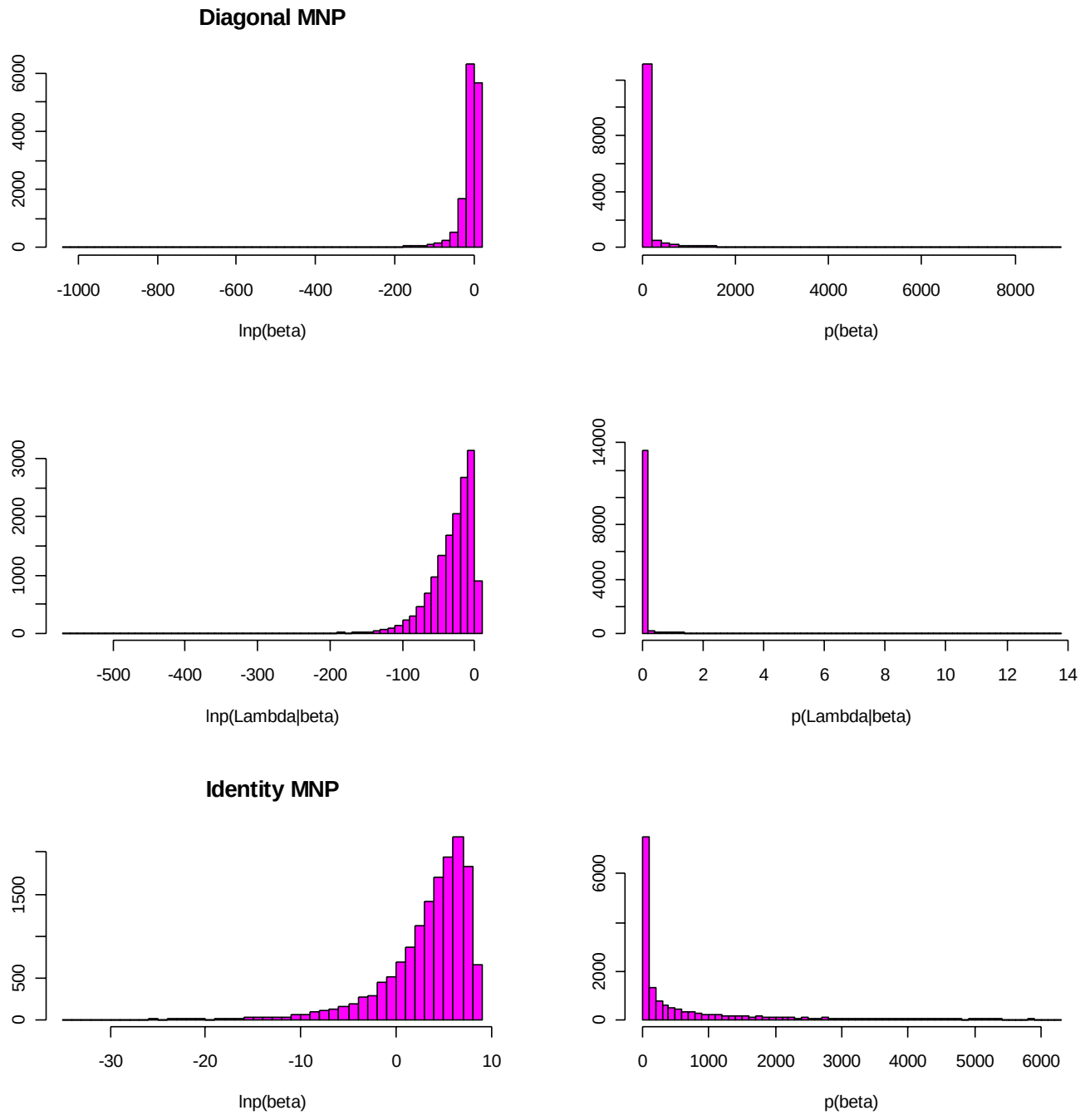


Figure 6.5
Posterior Distribution of Expected Revenues for Various Coupon Values