Bayesian Modeling in Finance

Xiaojing Dong[1]
Carrie H Pan

Leavey School of Business, Santa Clara University, Santa Clara, CA 95053

Published at *Journal of Investment Management* 2013 Vol. 11(1), pages 82-97

ABSTRACT

The Bayesian statistical method provides an alternative approach to study some of the classical problems in finance. In the existing finance literature, research that uses Bayesian econometrics is primarily in the area of asset pricing. Bayesian applications in corporate finance have been rather limited, despite its great potential as a viable alternative to address some challenging problems in corporate finance that are difficult to solve with the traditional approach. Bayesian estimation techniques, the Markov Chain Monte Carlo (MCMC) methods in particular, are very conductive to estimating non-linear models with high-dimensional integrals in the likelihood or models with a hierarchical structure. In this paper, we outline the basic concepts of Bayesian modeling, describe most commonly used estimation techniques, and review its applications in the existing finance literature.

[1] Dong's email address is xdong1@scu.edu and Pan's email is chpan@scu.edu. We are grateful to Sanjiv Das and John Heineke for their helpful comments and suggestions.

Introduction

Most quantitative methods and statistical analysis in finance take the classical (frequentist) approach, which essentially considers the probability of an event the limit of its long-run frequency. Data are assumed to be a repeatable random sample from the underlying population that has a distribution with unknown but fixed parameters. Inference is based on unbiased estimators, hypothesis testing and confidence intervals, and often relies on large-sample approximations. In reality, however, financial data violate these assumptions most of the time. For example, rare events such as financial crises do not repeat frequently enough, so large samples of time-series data are not available. Moreover, financial variables tend to behave differently during crisis. Stock market volatility increases during a crisis period, as do the correlations cross different stock markets (see e.g., Bekaert et al. 2005). In this case, the underlying parameters might be random variables themselves.

Some of these strong assumptions are not important, nor relevant in a Bayesian framework. Bayesian statistics considers data fixed and parameters uncertain. In the Bayesian framework, probability distributions are subject to change as new information (from data) becomes available. In this paper, we outline the basic concepts of Bayesian modeling, describe most commonly used estimation techniques, and review some of its applications in the existing finance literature.

Bayes' Theorem and Bayesian Model Analysis

Bayes' Theorem

Bayes' Theorem was first discovered in the 1740s by the English Reverend Thomas Bayes, when he was trying to learn how to infer causes from effects. It was eventually published after his death (Bayes 1763), but made little impact. Over a decade later, in 1774 Pierre-Simon Laplace, a French mathematician, independently rediscovered the theorem and published it (Laplace 1774). Laplace continued to use it, extended it, and made it more popularly known. Bayes theorem is simple and the idea is the following: suppose we denote E as the event or what actually happened,

and denote C_i as the i_{th} possible cause that could lead to the event E. The Bayes' theorem states that given the event has happened, the probability of each possible cause C_i is *proportional* to the product of two probabilities, (1) $P(E|C_i)$, the conditional probability of E given cause C_i; and (2) $P(C_i)$, the probability of cause i without knowing the event:

$$P(C_i|E) = \frac{P(E|C_i)P(C_i)}{\sum_j P(E|C_j)P(C_j)} \quad (1)$$

Traditionally, interest was in the probability of an event happening given the causes. But Bayes and Laplace were interested in making inferences about the cause, given the observed event. The Bayes' probability is therefore also referred to as a "reverse probability". Although simple, this theorem made a profound impact on the way statistical inferences are conducted. This can be demonstrated from Bayesian model analysis and its deviations from the classical approach.

Bayesian Model Analysis

In the context of modeling analysis, the objective is to obtain statistical inferences regarding model parameters for given data. Following conventional notation, we use θ to denote the parameters in the model, and y to denote the data. According to Bayes Theorem, we can write

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)} \quad (2)$$

In this equation, $p(\theta)$ denotes the *prior* distribution of the parameters, representing the analyst's knowledge about the parameters before seeing the data. $p(y|\theta)$ denotes the distribution of the data conditional on the parameters, the *likelihood* function. $p(y)$ denotes the marginal distribution of the data, and can be obtained by integrating the likelihood function over all possible θ, that is $p(y) = \int p(y|\theta)df(\theta)$. This is the continuous version of the denominator in equation (1). Since $p(y)$ does not involve θ, we can drop it from the above equation when inferring parameter θ for given data y and rewrite the equation as,

$$p(\theta|y) \propto p(y|\theta)p(\theta) \quad (3)$$

$p(\theta|y)$ on the left-hand side represents the (inferred) distribution of the parameters given the data, and it is called the *posterior* distribution of the parameters. The goal in Bayesian model analysis is to obtain the posterior distribution of the model parameters conditional on the data. Equation (3) represents the basic idea of the Bayes' Theorem, that is

Prior belief (*Prior*) + Data (*Likelihood*) → Updated belief (*Posterior*)

Therefore, the posterior distribution for parameters θ combines two sources of information, the likelihood function that captures information in the data, and the prior distribution that represents additional information available to the analyst.

In setting up the prior distribution, no information from the data is required. The information in the data is incorporated via the likelihood function. According to the *likelihood principle*, the likelihood function contains all relevant information from the data (Berger and Wolpert 1984). Bayesian modeling analysis is a likelihood-based approach, which has advantages over a non-likelihood-based method, as discussed in Kim et al. (2007). In addition, it has been shown that inference based on the likelihood approach reaches the Cramér-Rao lower bound, and is asymptotically efficient (Greene 2008, page 493).

There are two important points to note here. First, inferences of the parameters $p(\theta|y)$ are conditional on the data, which is different from classical statistical inference. The classical approach is based on sampling theory, where the data used in estimating the model are considered to be a random representation of the "population", which can be only achieved when the sample size is infinite. The statistical inferences obtained from such data tell us something about the population. In such statistical analysis, we care about the properties of statistical inferences when the sample size is infinite, which is referred to as the *asymptotic property*. In Bayesian approach, inference is conditional on the data, so there are no such concepts as "sampling" and "population". As a result, even when the size of the data is small, a Bayesian approach is still applicable. Second,

in the classical approach, we obtain point estimates of the model parameters, with asymptotic properties. When using these model estimates for forecasting, it is sometimes tedious to obtain the confidence intervals of the predicted dependent variables. In the Bayesian approach, however, given that the model result is the *distribution* of the model parameters conditional on the data, we can simulate the parameters from the posterior distributions, plug in the model and obtain the simulated distribution of the dependent variable. This approach has proved to be able to solve interesting yet important problems that might be difficult using a classical approach (Rossi et al. 1996, Dong 2007). In this case, the distribution of the dependent variable can be achieved through Monte Carlo simulation. It is not necessary to derive the asymptotic properties of the predicted variables, using some approximation method, such as the Delta method (Cameron and Trivedi 2005, page 231), as maybe required by a classical approach.

Furthermore, the incorporation of the prior in Bayesian inference diverges sharply from the classical statistical inference.[2] In Bayesian model analysis, getting prior knowledge of the model parameters is an important step. This is because all additional knowledge regarding the parameters other than the data can be incorporated into the modeling analysis through the prior. In Bayesian modeling, prior belief is represented by the prior distribution of the model parameters, which could be formulated in many different ways. The flexibility of prior distributions provides the analyst with a formal way to incorporate any information, in addition to the data and model, into the process of parameter inferences. This information could be obtained from a different dataset (Dong 2007, Shin et al. 2012), from prior knowledge (Allenby et al. 1995), or from economic theory (Montgomery and Rossi 1999). Sometimes prior information about the parameters is too limited for the analyst to specify the prior distribution. In this case, a non-informative prior in the form of a diffuse distribution could be deployed. It is then necessary to ensure that such a distribution

[2] The emphasis on prior belief and the idea of subjective probability in Bayesian analysis have drawn criticism from the frequentists. A persuasive proponent is a quote attributed to a renowned statistician I. J. Good, "*the subjectivist (i.e. Bayesian) states his judgments, whereas the objectivist sweeps them under the carpet by calling assumptions knowledge, and he basks in the glorious objectivity of science*."

assumption is appropriate in the modeling context, and the analyst needs to understand its impact on the posterior distribution of the parameters (see for example Jacquier et al. 2004). It could also be obtained from another model in a hierarchical structure.

Hierarchical Bayesian Models

To illustrate the structure of Hierarchical Bayesian models, consider the following regression model estimated with a panel dataset, where *i* indexes for firm and *t* for time. The setup of the model is similar to a regular OLS model, where the error term ϵ_{it} is assumed to be i.i.d. and follow a normal distribution $N(0, \sigma^2)$. In contrast to the OLS method, the HB model accounts for unobserved heterogeneity by allowing each *i* have its own set of parameters β_i.

$$y_{it} = X_{it}\beta_i + \epsilon_{it}, \qquad where\ \epsilon_{it} \sim N(0, \sigma^2) \tag{4}$$

In this case, the prior distribution of each β_i for all *i* can be assumed to follow a normal[3] distribution, such as

$$\beta_i \sim N(\bar{\beta}, \Sigma)$$

The values of the parameters in the prior distribution, i.e. mean $\bar{\beta}$ and covariance matrix Σ, are hard to come by, so they are represented with distributions. These distributions form another layer of prior distributions, which makes the model a hierarchical model. These parameters mean $\bar{\beta}$ and covariance matrix Σ are called *hyperparameters*. As an example, we can assume that they are independent, and the prior for $\bar{\beta}$ follows a normal distribution,

$$\bar{\beta} \sim N(\beta_0, A^{-1})$$

[3] For the purpose of demonstration here, we use a Normal distribution assumption for the prior. Such a prior is conjugate and will lead to convenience in estimation (see the discussion in the session 'Conjugate Priors'). The prior distribution of β_i could be assumed to other distribution function as well.

The values of β_0 and A^{-1} are specified by the analyst, based on the knowledge of the analyst. If no information is available, a diffuse prior could be used. We can set $\beta_0 = 0$, and $A = 0.01 * I$, where I is an identity matrix.

The prior for the covariance matrix Σ is assumed to follow an Inverse-Wishart (IW) distribution, with

$$\Sigma \sim \text{IW}(n_0, V_0)$$

An Inverse-Wishart distribution is the multi-dimensional Inverse-Gamma (IG) distribution, and is specified with two parameters: the degree of freedom n_0 and the scale matrix V_0. Denote p as the dimension of matrix V_0[4]. For the distribution to be valid, the degree of freedom parameter n_0 needs to satisfy $n_0 > p - 1$, and V_0 needs to be positive definite. The distribution has a conjugate prior, and is commonly used as the distribution function for the covariance matrix of a multivariate normal distribution. In Bayesian estimation, the values of n_0 and V_0 need to be specified. If a diffuse prior is desired, n_0 should be small, but also satisfying the condition $n_0 > p - 1$. After choosing n_0, the analyst can specify the scale matrix V_0 based on the fact that the mean of the Inverse-Wishart distribution is

$$\frac{V_0}{n_0 - p - 1}$$

For more discussion regarding the properties and implementations of Inverse-Wishart distribution, interested readers are referred to a Bayesian text book, such as Gelman et al. (1995) page 474.

In this example, the goal of model estimation is to obtain firm level parameter estimates β_i for all i. Its prior distribution $\beta_i \sim N(\bar{\beta}, \Sigma)$ is the same for all i. This normal distribution can be considered as the population level distribution across all i. This population level distribution is achieved based on the inference of the firm-level estimates β_i. In most cases, each i may have only a limited number of data points, which subjects the inference to various sources of noise. In the

[4] That is, if V_0 is a 4x4 matrix, then p=4.

Bayesian inference, the prior distribution $\beta_i \sim N(\bar{\beta}, \Sigma)$ is effectively combined with the data concerning each firm *i* to achieve the posterior distribution of β_i for any particular *i*. In other words, this Hierarchical Bayesian framework "borrows" information from all other *i*'s to obtain the posterior inference about a particular *i*.

As such, a Hierarchical Bayesian model provides an effective way of obtaining individual-level estimates. For example, tests of asset pricing models usually take a portfolio-based approach to reduce the errors-in-variables problems of estimated betas (see e.g., Blume 1970; Black et al. 1972; Fama and MacBeth 1973). However, many recent studies point out the shortcomings of the portfolio-based approach such as data snooping, test efficiency, or difficulty in selecting the appropriate test portfolios (see e.g., Lo and MacKinlay 1990; Ahn et al. 2009; Ang et al. 2010; Lewellen et al. 2010). Cederburg et al. (2011) develop a Hierarchical Bayesian model to test the Capital Asset Pricing Model (CAPM) at the firm-level. Their model has the following structure:

$$r_{ity} = \alpha_{iy} + r_{mty}\beta_{iy} + \epsilon_{ity}, \qquad \epsilon_{ity} \sim N(0, \sigma_{iy}^2) \tag{5}$$

$$\alpha_{iy} = X_{iy}\delta_y + u_{iy}, \quad u_{iy} \sim N(0, \sigma_{\alpha y}^2) \tag{6}$$

$$\delta_y = \overline{\delta} + v_y, \quad v_y \sim MVN(0, V) \tag{7}$$

Where r_{it} denotes the excess return on stock *i* in subperiod *t* over time period *y*, r_{mt} denotes the excess market return, and X_{iy} is a matrix including a constant and firm characteristics measured at the beginning of period *y*. The unobserved firm-level heterogeneity is addressed through firm-specific parameters, α_{iy} and β_{iy}. The firm-specific α_{iy} is assumed to have a hierarchical structure. Each α_{iy} is modeled as a function of firm characteristics, as shown in equation (4). The dependence of α_{iy} on firm characteristics can also vary across different time periods, as captured by δ_y. Cederburg et al. (2011) examine nine CAPM anomalies, including size, book-to-market, momentum, reversal, profitability, asset growth, net stock issues, accruals, and financial distress,

over the period of 1963-2008. Their results with firm-specific alphas suggest that much of the evidence using the portfolio-based approach against the CAPM is overstated.

Others have estimated more complicated hierarchical models. For example, Greyserman et al. (2006) estimate a hierarchical model for (μ, Σ) in portfolio optimizations (Markowitz 1952). They compare the performance of portfolios constructed weights estimated from a classical mean-variance optimization model with no hierarchical structure at all, a partial hierarchical model in which μ has a hierarchical structure (a shrinkage estimator for the mean), and a full hierarchical model with hierarchical structures for both μ and Σ. Greyserman et al. show that portfolios constructed with the full hierarchical model outperform those constructed with other models. Young and Lenk (1998) estimate a hierarchical multifactor model in which the alphas, betas, and variances are linear functions of firm characteristics. They show that the use of cross-sectional information in the Hierarchical Bayesian models leads to smaller estimation errors of the factor model parameters at the firm-level, and the improvement in estimation accuracy further leads to improved portfolio performance.

The fact that these hierarchical models have multiple levels provides a natural way of grouping the parameters, which makes it directly applicable to Gibbs sampling and other Markov Chain Monte Carlo (MCMC) methods, which will be discussed later. Before presenting the estimation approaches using MCMC methods, we discuss Conjugate Priors, another very important concept in Bayesian model analysis in the next section.

Conjugate Priors

In some case, a particular distribution assumption on the priors could allow the analyst to conveniently obtain the posterior distribution of the model parameters. Conjugate priors is one example. When the posterior derived from a prior and likelihood is in the same class of distribution as the likelihood, the prior is called a *conjugate* prior. When a conjugate prior exists, the distribution

function of the posterior is known, and the parameters in the posterior distributions can be derived analytically.

Unfortunately, most likelihood functions do not have a conjugate prior, except likelihood functions in the exponential family of distributions.[5] The exponential family consists of many commonly used distributions, especially those with an exponential term in their probability distribution function (PDF). Such distributions include Normal, Exponential, Gamma, Wishart, Binomial, Bernoulli, Poisson, Negative Binomial, Dirichlet, etc. However, not all of these distributions have standard conjugate priors (Consonni and Veronese 1992). In fact, Poisson likelihood does not have a conjugate prior. But if the parameter of a Poisson distribution follows a Gamma distribution, it becomes a Negative Binomial (or Poisson-Gamma) distribution, for which conjugate priors exist.

When a conjugate prior exists for a distribution in the exponential family, the product of the prior and the likelihood is a multiplication of two exponential terms, which can be rewritten as the exponential of the sum of the two parameters (one from the prior and the other from the likelihood). The sum then becomes the parameter of the posterior distribution.

For example, consider a dataset with N observations $x_1, x_2, \dots, x_N$, and assume that they are generated from a normal distribution with known variance σ^2. We want to apply the Bayesian approach to obtain estimates for the mean μ. The likelihood function of the data conditional on the unknown parameter μ and the known parameter σ^2 can be written as

$$p(data|\mu,\sigma^2) \propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

We use the proportional operator $\propto$ here, as the scale parameter (functions of π and σ^2) is omitted for simplification purposes. We use the conjugate prior for the mean μ, that is

[5] For the detailed discussion about the conjugacy of the exponential family, interested readers are referred to Section 5.2 in Bernardo and Smith (1994) and Section 2.3 in Gamerman and Lopes (2006).

$\mu \sim N(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are the mean and variance of the prior distribution and their values are chosen by the analyst. The PDF for the prior distribution is

$$p(\mu) \propto \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)$$

The posterior distribution for μ conditional on the data and the known parameter σ^2, $p(\mu|data, \sigma^2)$, can be obtained through the multiplication of the two normal distribution functions, that is

$$p(\mu|data, \sigma^2) \propto \exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2\right)\prod_{i=1}^{N}\exp\left(-\frac{1}{2}\left(\frac{x_i-\mu}{\sigma}\right)^2\right)$$

Combining the two exponential functions, the right hand side becomes

$$\exp\left(-\frac{1}{2}\left(\frac{\mu-\mu_0}{\sigma_0}\right)^2-\frac{1}{2}\left(\frac{\sum_{i=1}^{N}(x_i-\mu)}{\sigma}\right)^2\right)$$

$$= \exp\left(-\frac{1}{2}\left(\mu^2\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)-2\mu\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)\right)+C_1\right)$$

$$= \exp\left(-\frac{1}{2}\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)\left(\mu-\frac{\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)}\right)^2+C_2\right)$$

In the above equations, C_1 and C_2 are constants relative to μ. Comparing the last equation with a Normal distribution PDF $\exp\left(-\frac{1}{2}\frac{1}{var}(\mu-mean)^2\right)$, we can obtain the parameters of the posterior distribution for $p(\mu|data, \sigma^2) \sim N(mean, var)$, where

$$var = \left(\frac{1}{\sigma_0^2}+\frac{N}{\sigma^2}\right)^{-1}$$

$$mean = \left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^{N}x_i}{\sigma^2}\right)\times var$$

The existence of a conjugate prior allows the analyst to arrive at the posterior distribution analytically, as demonstrated above. When a conjugate prior does not exist, the distribution of the posterior is unknown. One has to resort to the MCMC method to simulate the posterior distribution.

MCMC and Data Augmentation

The Markov Chain Monte Carlo (MCMC) method provides a way to simulate from a distribution using a Markov Chain. In layman's terms, a Markov Chain simulation refers to a process of stochastically generating a series of numbers, where the successive number is generated probabilistically based only on the preceding one and is not influenced by the path it took to achieve the preceding quantity. One important property for a Markov Chain is that when the chain is long enough, the limiting quantities will reach a stationary distribution[6]. The idea of MCMC is to create a Markov Chain so that its stationary distribution is the posterior distribution. Once the posterior distribution has been simulated, the analyst can obtain any descriptive statistics about the distribution. That includes but is not limited to (1) mean of the posterior distribution, which corresponds to the point estimate in the frequentist approach; (2) standard deviation, which corresponds to the standard errors in the frequentist approach; and (3) confidence intervals. In practice, it is achieved by generating a really long Markov Chain, even after convergence is attained. The simulated numbers at the beginning, which is called the "burn-in" period, are discarded. The analyst can then obtain descriptive statistics from the many draws when the chain converges after the burn-in period.

MCMC starts from an initial value, and at every step the next step is defined based on the current value and the transition probability function. The key challenge in this process is therefore to define the transition probabilities for the Markov Chain, so that the chain will converge to the posterior distribution. Two widely adopted approaches are employed to achieve this goal, Gibbs sampling and the Metropolis-Hastings algorithm. Another important approach that is directly

[6] For more details about Markov Chain and its properties that are relevant to Bayesian analysis, interested readers are referred to Chapter 4 of Robert and Casella (1999).

related is the data augmentation approach, which dramatically simplifies the estimation process in many Bayesian model analyses. In the following, we present the Gibbs sampling approach and the Data augmentation technique, followed by a discussion of the Metropolis-Hastings Algorithm.

Gibbs Sampling

In most modeling analysis, the posterior distribution of the model parameters is multi-dimensional. Simulation from a multi-dimensional distribution could be formidable, if possible at all. Gibbs sampling provides an approach to attain the high-dimensional joint distribution through sequentially simulating from a series of conditional distributions. Gibbs sampling is named after the physicist Josiah W. Gibbs (1839-1903), known for his major contributions to statistical physics. This approach was developed by Geman and Geman (1984), about eight decades after the death of Gibbs. The concept of Gibbs sampling is detailed in the following:

Denote θ as a vector containing all the model parameters,

$$\theta = \{\theta_1, \theta_2, \dots, \theta_g\}$$

$\theta_i, i = 1,2, \dots, g$, could represent one parameter or a set of parameters (vector). In some cases, the model parameters are naturally grouped based on the specifications of the model. For example, in Hierarchical Bayesian models, some parameters are individual-level parameters, whereas others are at the population-level. As another example, some studies control for endogeneity by simultaneously estimating multiple models (see for example Yang et al. 2003 and Dong et al. 2009). The parameters associated with each model can be put in the same group. When such a grouping is not well defined by the model, the analyst is free to group the parameters as she wishes. If too many groups are defined, it takes a long time before all the parameters can be simulated once in the Gibbs sampler, which will slow down the estimation. However, if too few groups are created and in each group there are a large number of parameters, the estimation might still have to face a

high dimensional problem that the Gibbs sampler was set to resolve. Once the groups was defined balancing the tradeoffs, the Gibbs sampler can be developed.

Let $\pi(\theta)$ denote the joint posterior distribution of all the parameters, which is what the Bayesian modeling analysis sets to achieve. According to the Gibbs sampling approach, to obtain $\pi(\theta)$, one can simulate each θ_i from the full conditional distribution of θ_i, conditional on the rest of parameters. That is,

Step 1: Simulate θ_1 from the conditional distribution $f_1(\theta_1|\theta_2, \dots, \theta_g)$, get a value of θ_1, denoted as θ_1^t.

Step 2: Simulate θ_2 from the conditional distribution $f_2(\theta_2|\theta_1^t, \theta_3, \dots, \theta_g)$. Note that the conditional distribution is conditional on the current value of θ_1^t obtained from the last step. From this step, we get θ_2^t.

Step 3: Simulate θ_3 from the conditional distribution $f_3(\theta_3|\theta_1^t, \theta_2^t, \dots, \theta_g)$ using the current values of θ_1^t and θ_2^t, and get the new value θ_3^t.

Continue with the above steps, draw each θ_i conditional on the current values of the rest, until the chain reaches convergence.

To see how Gibbs sampling is applied, consider again the model in Cederburg et al. (2011). The model parameters that need to be estimated are $\{\alpha_{iy}, \beta_{iy}, \sigma_{iy}^2, \delta_y, \overline{\delta}, \sigma_{\alpha y}^2, V\}$. In this study, diffuse priors are adopted. The prior for $\overline{\delta}$ is assumed to be multivariate normal with mean zero and really large variances, $\overline{\delta} \sim MVN(0, 100\boldsymbol{I})$, where $\boldsymbol{I}$ represents the identity matrix. The prior for firm-level betas β_{iy} is assumed to follow a normal distribution with a mean of one and a large variance, $\beta_{iy} \sim N(1, 10)$. In both cases the prior has a large variance so that the prior mean would have little impact on the posterior. The prior distributions for $\{\sigma_{iy}^2\}$ and $\{\sigma_{\alpha y}^2\}$ are specified as Inverse-Gamma (IG), and the prior for V is assumed to be Inverse-Wishart (IW), which is the multidimensional version of the Inverse Gamma distribution. As discussed earlier, these are all conjugate priors for

the model parameters. With these prior distribution assumptions, Cederburg et al. use a Gibbs sampler to draw from the full conditional posterior distributions for all parameters:

Step 1: Draw $\alpha_{iy}, \beta_{iy} | \sigma_{iy}^2, \delta_y, \sigma_{\alpha y}^2 \sim N(\,)$ for each stock *i* in each year *y*;

Step 2: Draw $\sigma_{iy}^2 | \alpha_{iy}, \beta_{iy} \sim IG(\,)$ for stock *i* in each year *y*;

Step 3: Draw $\delta_y \left| \{\alpha_{iy}\}, \beta_{iy}, \overline{\delta}, V \sim N(\,) \right.$ for each year *y*;

Step 4: Draw $\sigma_{\alpha y}^2 \left| \{\alpha_{iy}\}, \delta_y \sim IG(\,) \right.$ for each year *y*;

Step 5: Draw $V | \{\delta_y\} \sim IW(\,)$;

Step 6: Draw $\overline{\delta} | \delta_y \sim N(\,)$.

Each iteration of these six steps forms one draw from the joint posterior distribution of all the parameters. These steps are repeated until the MCMC converges. The additional draws after reaching the convergence are used to obtain the descriptive statistics of the posterior distributions of the model parameters.

Data Augmentation

The Data Augmentation approach was developed by Tanner and Wong (1987), and has been widely adopted in solving Bayesian models. According to Tanner and Wong (1987), the data augmentation method can be applied whenever augmented data (a) are easier to be analyzed; and (b) make the MCMC process easier to generate given the model parameters. In other words, it is widely applicable, as long as it provides convenience in generating the MCMC. The basic idea of data augmentation is to facilitate the MCMC process by simulating some additional random variables that are not model parameters.

For example, Albert and Chib (1993) developed a Gibbs sampler with data augmentation to facilitate the estimation of a Probit model. The likelihood function of a Probit model involves integration over a normal distribution, which does not have a closed form solution. Therefore it has been the major obstacle to using a likelihood-based approach (such as Bayesian inference) in

estimating a Probit model. The main difficulty in the integration happens when trying to connect the discrete values on the left hand side with the continuous ones on the right hand side in a Probit model. To break the deadlock, Albert and Chib (1993) suggest augment the data by simulating an additional continuous variable that connects both sides of the model. For example, in a Binary Probit model:

$$y_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim N(0,1), \ \ i = 1,2,\dots,P$$

$$Y_i = \begin{cases} 1, & if\ y_i > 0 \\ 0, & otherwise \end{cases}$$

In this model y_i is latent, and not observed. In estimating the model with data augmentation, its value is augmented, which dramatically simplifies the estimation process, In particular, the Gibbs sampler is:

Step 1: Draw β conditional on the latent values y and the observed data X. This step is similar to estimating a normal regression with conjugate priors for the β. Denote the prior as $\beta \sim N(\beta_0, A^{-1})$, the full conditional distribution (the posterior distribution of β conditional on the latent values y_i) follows a normal distribution with

$$\mu_\beta = (X'X + A)^{-1}(X'y + A\beta_0)$$

$$\Omega_\beta = (X'X + A)^{-1}$$

Step 2: Draw the latent values $y_i, i = 1,2,\dots,P$ conditional on β and the observed data X. P independent draws of y_i are simulated. Each draw is simulated from a truncated normal distribution $TN(X_i\beta, 1)$. If $Y_i = 1$, the truncation is $(0, \infty)$; otherwise, the truncation is $(-\infty, 0)$.

Besides simplifying the modeling analysis, data augmentation can also be useful when dealing with a very common problem in finance - missing data. Korteweg (2011) notes that even key variables such as total assets, capital expenditures, and market leverage are missing in Compustat at alarmingly high rates. Some analysts exclude observations with missing data, others replace the missing data with some arbitrary values such as sample median or industry median. Neither seems to be an ideal solution. Dropping observations with missing data throws away all the

information in other variables associated with these observations. Filling missing values with a point estimate ignores the variation in the filled-in data. Data augmentation offers a convenient way to address this problem by treating the missing observations as latent values (y^*) that are added to the estimation. The joint distribution of the missing values and the model parameters can be derived, which demonstrates one of the convenient properties of Bayesian inference, that is, data and parameters are treated as the same. The joint distribution of the parameters can be obtained by margining out the parameters. All these steps are natural in Bayesian estimation using MCMC.

Korteweg (2010) uses this technique to handle missing corporate bond values in his study of the net benefit of leverage. He models net benefit of debt relative to total firm value as a function of firm characteristics, interactions of firm characteristics and leverage, and interactions of firm characteristics and squared leverage. His model contains a firm value equation, a beta equation, one-factor return equations for equity and debt, and an AR(1) equation for unlevered beta. To handle missing data in bond returns for bonds that are traded infrequently, Korteweg treats missing values as additional model parameters. He finds that net benefits of leverage accounts for as much as 5.5% of firm value, and firms are on average underlevered relative to the optimal capital structure.

Metropolis-Hastings Algorithm

Gibbs sampling allows the analyst to decompose a complicated high dimensional problem into a series of smaller models that are easier to solve. The simpler models still need to be solved. If a conjugate prior exists, the posterior distribution can be derived analytically, and the Gibbs sampler can fulfill the task. However, if a conjugate prior does not exist, the estimation process becomes difficult. In fact, besides higher requirements on computational power, not being able to solve the models without a conjugate prior was another major hurdle that prevented Bayesian analysis from being widely accepted until 200 years after the Bayes Theorem was discovered.

In 1953, a group of physicists from Los Almos proposed a simulation method for calculating integrations (Metropolis et al. 1953). Over a decade later, this method was extended to more general cases (Hastings 1970), and is therefore recognized as the Metropolis-Hastings (MH) algorithm. As one of the MCMC approaches, the MH algorithm provides a way to obtain transition probabilities so that the chain will converge to the posterior distribution of the model parameters.

In the MH algorithm, the transition probabilities are unknown, so the approach allows the analyst to use a proposal transition function, denoted as $q()$. The Markov Chain can be simulated with a starting value θ^0 and the proposal transition function $()$. Given that $q()$ is not the true transition matrix that will converge to the posterior distribution as a stationary distribution, some adjustment (correction) is necessary. The adjustment is to decide the probability of accepting the proposed value generated from the proposal transition function, denoted as $\widetilde{\theta^t}$. The acceptance probability is denoted as α, and is calculated as

$$\alpha = \min\left(\frac{\pi(\widetilde{\theta^t})q(\theta^{t-1})}{\pi(\theta^{t-1})q(\widetilde{\theta^t})}, 1\right) \tag{8}$$

That is, with probability α the chain accepts the proposed value and moves on with the chain $\theta^t = \widetilde{\theta^t}$; and with probability $1-\alpha$, the chain rejects the proposed value and repeats the current value, that is $\theta^t = \theta^{t-1}$.

The MH algorithm is a very powerful approach and widely applicable, as it does not require a functional form of the parameter posterior distribution $\pi(\theta)$. The development of this method helped fuel the wider acceptance of Bayesian methods in estimation. In addition, given that the calculation of the acceptance probability α involves the ratio of $\pi(\theta)$ at two different θ values, the MH algorithm is still applicable even when the posterior distribution is only known up to a scale. This feature is especially attractive to Bayesian analysis, as it dramatically simplifies the calculation of the posterior distribution $\pi(\theta)$. To understand this, recall that equation (3) shows that the posterior distribution is *proportional* to the product of the prior and likelihood function. In most

cases of Bayesian modeling, getting the scale parameter of the posterior distribution is cumbersome. To achieve that, one needs to calculate the integral of the unscaled posterior distribution, which was obtained through a multiplication of two probability distribution functions. The MH algorithm allows the analyst to skip such a step, which makes it well accepted and suitable to most Bayesian modeling analysis[7].

To better understand the MH algorithm, let's compare it with the Acceptance-Rejection (AR) sampling approach[8]. The concepts of these two approaches are quite similar, in that a proposal distribution is required in both approaches. After simulating a "proposal", the AR approach will reject the proposed value if it is too big; while the MH will *keep* the proposed value if "correction" is needed. Another main difference between these two approaches is that in the AR approach, the ideal proposal distribution is larger than or the same as the desired distribution function on the support of the distribution, while this is not necessary in the MH approach. In fact, the MH algorithm does not put any restrictions on the proposal function $q()$, nor any theoretical guidance on what might make a good $q()$ in order to result in good convergence. In practice, the most commonly used proposal function is a normal distribution centered at the current value θ^t, and variance s^2 chosen by the analyst. Such a MH algorithm is also called a Random-Walk MH algorithm, which is specified as follows.

Step 0: Start with θ_0,

Step 1: draw the proposal $\widetilde{\theta^t} \sim N(\theta^{t-1}, s^2)$

Step 2: calculate the acceptance probability α by evaluating equation (4).

Step 3: decide whether to accept the proposal. With probability α, $\theta^t = \widetilde{\theta^t}$; and with probability $1-\alpha$, $\theta^t = \theta^{t-1}$

7 For more detailed discussion on the MH algorithm and its applications as well as MCMC methods in general, the readers are referred to Gilks et al. (1996).

8 For a detailed discussion on the AR approach, please refer to Law and Kelton (2000), pages 452-458.

One of the key parameters in this process is the variance of the proposed normal distribution s^2. If the s^2 is too big, many of the proposed values will be rejected, and the chain will be moving too slow. On the other hand, if it is too small, even though most of the proposed values are accepted, they may be very close to each other so that the chain navigates the space too slowly to reach the convergence. By adjusting the value of s^2, one can balance the trade-offs. A rule of thumb is that an acceptance rate around 30% is a good number.

Bayesian estimation of time-varying volatility models such as GARCH relies on Metropolis-Hastings algorithm because the likelihood function in such a model does not have a conjugate prior, and it is impossible to draw directly from the posterior distribution of the parameters (see for example Chib and Greenberg 1994; Müller and Pole 1998; Nakatsuma 2000). Consider the GARCH (p, q) model (Bollerslev 1986):

$$\varepsilon_t = z_t \sigma_t, \qquad z_t \sim N(0,1)$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

Where $\sigma_t^2$is the conditional variance of $\{\varepsilon_t\}$ at time $t, t = 0, 1, \dots, T$; p, q are integers where $p > 0$, $q \geq 0, \alpha_0 > 0,\ \alpha_i \geq 0, i = 1, \dots, p, and\ \beta_j \geq 0, j = 1, \dots, q$; and $\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j < 1$ to ensure stantionarity. The parameter vector $\{\alpha_0, \alpha_1, \dots, \alpha_p, \beta_1, \dots, \beta_q, \sigma_0^2\}$ for a GARCH (p, q) model is not of high dimension so that the repeat rate is reasonable in MH draws and convergence is fairly quick.

Vrontos et al. (2000) use a random walk MH algorithm with an incremental normal density $N(0, \sigma^2)$ to estimate several GARCH and EGARCH models. The candidate draw is made by adding a random value generated from $N(0, \sigma^2)$ to the current value, and σ is set to generate no more than 50% repeats. In addition, they use a reversible-jump MCMC algorithm to estimate several models simultaneously. In addition to the posterior densities for each model, this algorithm also generates the posterior probabilities for each model, which can be used to derive posterior

estimates for the optimal composite model by averaging the posteriors of each candidate model with weights being the corresponding posterior model probabilities. Their approach therefore accounts for model uncertainly in volatility forecasting.

The Metropolis-Hastings algorithm has been essential for Bayesian estimation of many stochastic volatility models (see for example Jacquier et al. 1994, 2004; Kim et al. 1998; Eraker et al. 2003; and Eraker 2004). For a detailed review of the literature and how to implement Bayesian estimation of stochastic volatility models, please consult Jacquier and Polson (2010).

The Gibbs sampling and Metropolis-Hastings algorithm can be used either separately as discussed above, or together in a hybrid form. With Gibbs sampling the analyst can first decompose a complicated model into a sequence of simpler models. If each of these decomposed models has a conjugate prior, simulating from the full conditional distribution is straightforward. If in some steps of the Gibbs sampler, a closed-form conditional posterior distribution does not exist, a Metropolis-Hastings approach is necessary. *MH within Gibbs* algorithm is therefore widely applicable in solving almost all Bayesian models.

For example, Chib and Winkelmann (2001) model correlations among count data using correlated latent effects. It is represented by correlated random terms additive to Poisson parameters in a log-link specification[9]. These additive error terms are assumed to follow multivariate normal distribution, which captures the correlation among these count data using the correlations among the normal error terms. The resultant model is a hybrid of Poisson and Log-Normal. The natural grouping of the model parameters are therefore based on which model (Poisson or Log-Normal) these parameters are associated with. For the Gibbs step associated with the Poisson model, no conjugate prior exists, therefore an MH algorithm is employed.

Besides the Poisson model, another commonly used model without a conjugate prior is the Logit choice model. The logit model is commonly used to model consumer's choices, and it is

[9] That is, the log of the Poisson parameter is specified as a linear function.

especially popular in Marketing literature since the seminal paper by Guadagni and Little (1983). When unobserved heterogeneity is introduced into the model and estimated with Hierarchical Bayesian approach, the MH within Gibbs can be adopted (see for example Rossi et al. 1996; Allenby et al. 1998).

Summary

The Bayesian statistical approach provides an alternative way to study many classical problems in finance. Bayesian estimation techniques, Markov Chain Monte Carlo (MCMC) methods in particular, are very attractive to estimating non-linear models with high-dimensional integrals in the likelihood or models with a hierarchical structure. In many cases, the classical approach would have been difficult or infeasible in solving these problems. These include:

1. The Hierarchical Bayesian approach is natural for obtaining *unobserved heterogeneity* by estimating individual-specific parameters, and these parameters can be also specified as a function of individual characteristics (see for example Young and Lenk 1998; Greyserman et al. 2006; Cederburg et al. 2011).
2. The solution from the Bayesian approach is the joint posterior distribution of all the model parameters. In this case, *the stochasticity of the model parameters can be fully taken into consideration* when using MCMC outputs in an optimization or forecasting process. (see for example Allenby and Lenk 1994);
3. Bayesian approach is a likelihood-based approach, which, according to the likelihood principle, can achieve *efficiency*. However, the traditional maximum likelihood method cannot handle some complex financial models such as stochastic volatility models, which are non-linear and volatility is a latent variable. MCMC method provides a way to *solve such complicated models via simulation methods* (see for example Jacquier et al. 1994, 2004; Kim et al. 1998; Eraker et al. 2003 and Eraker 2004);

4. *Missing data* is a common problem in modeling analysis in corporate finance. The MCMC approach offers a convenient solution by including the missing observations as a latent variable into the model (see for example Korteweg 2010);

5. *Rare events* such as financial crises are difficult to study due to the lack of time series data. Similarly, it is challenging to investigate hedge fund returns with large tail risks or returns on assets that are traded infrequently. Sometimes the timing of the observed data is endogenous, leading to a *dynamic sample selection problem*. Korteweg and Sorensen (2010) examine venture capital investments in start-up companies using a dynamic sample selection model estimated using the Gibbs sampling approach.

In the existing finance literature, research that uses Bayesian econometrics is primarily in the area of asset pricing. Bayesian applications in corporate finance have been rather limited, despite its great potential as a viable alternative to address some challenging problems in corporate finance that are difficult to solve with the traditional approach. Bayesian methods and the MCMC approach are especially attractive in estimating non-linear models with high-dimensional integrals in the likelihood. Hierarchical Bayesian models deserve special attention. Corporate finance research often uses panel data, and the hierarchical structure of these models can incorporate many layers of heterogeneity, such as industry, firm, year, etc. A full discussion of this topic is beyond the scope of this primer, and we refer interested readers to Korteweg (2011) for a detailed discussion on the application of Bayesian econometrics in corporate finance.

Reference:

Ahn, D., J. Conrad, and R.B. Dittmar (2009), "Basis Assets," *Review of Financial Studies* 22, 5133-5174.

Aït-Sahalia, Y. and R. Kimmel (2007), "Maximum Likelihood Estimation of Stochastic Volatility Models," *Journal of Financial Economics* 83, 413-52.

Albert, J. H. and Chib, S. (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669–679.

Allenby, Greg M, Neeraj Arora, and James. L. Ginter (1995), "Incorporating prior knowledge into the analysis of conjoint studies," *Journal of marketing Research* 32, 152-162.

Allenby, Greg M, Neeraj Arora, and James. L. Ginter (1998), "On the Heterogeneity of Demand," *Journal of marketing Research* 35, 383-389.

Allenby, Greg. M. and Peter Lenk (1994), "Modeling household purchase behavior with logistic normal regression," *Journal of the American Statistical Association*, 89 (428), 1218-1231.

Ang, A., J. Liu, and K. Schwarz (2010), "Using Stocks or Portfolios in Tests of Factor Models," Working paper, Columbia Univeristy.

Bayes, Thomas (1763), "An Essay towards Solving a Problem in the Doctrine of Chances", *Philosophical Transactions of the Royal Society*, 53, pages 370-418. http://rstl.royalsocietypublishing.org/content/53/37 (accessed on 7/25/2012).

Bernardo, José-Miguel and Adrian F. M. Smith (1994), *Bayesian Theory*. Chichester: John Wiley & Sons, Ltd.

Bekaert, G., C. R. Harvey, and A. Ng (2005), "Market Integration and Contagion", *Journal of Business* 78, 39-69.

Berger, James O. and Robert L. Wolpert (1984), *The Likelihood Principle*, IMS Lecture Notes – Monograph Series. Hayward, CA: Institute of Mathematical Statistics.

Black, F., M. Jensen, and M. Scholes (1972), "The Capital Asset Pricing Model: Some Empirical Tests," In Jensen, M. ed., *Studies in the Theory of Capital Markets*, Praeger, New York, 79-121.

Blume, M. E. (1970), "Portfolio Theory: A Step toward Its Practical Application," *Journal of Business* 43, 152-173

Bollerslev, T. (1986), "Generalized autoregressive conditional heteroskedasicity," *Journal of Econometrics*, 31 (3), 307-327

Cameron, A. Colin and Pravin K. Trivedi (2005), *Microeconometrics: Methods and Applications*, Cambridge University Press: New York, NY.

Cederburg, S., P. Davies, and M. O'Doherty (2011), "Asset-Pricing Anomalies at the Firm Level," Working paper.

Consonni, G. and P. Veronese (1992), "Conjugate Priors for Exponential Families Having Quadratic Variance Functions," *Journal of the American Statistical Association* 87 (420), 1123-1127.

Chib, S. and E. Greenberg (1994), "Bayes Inference for Regression Models with ARMA(p,q) Errors," *Journal of Econometrics* 64, 183-206.

Chib, S. and R. Winkelmann (2001), "Markov Chain Monte Carlo Analysis of Correlated Count Data", *Journal of Business and Economic Statistics*, 19 (4), 428-435.

Dong, X. (2007), *Hierarchical Bayesian Method in the Study of Individual Level Behavior- in the context of discrete choice modeling with revealed and stated preference data*, VDM Verlag Dr. Mueller e.K.

Dong, X., P. Manchanda and P. K. Chintagunta (2009), "Quantifying the Benefits of Individual Level Targeting in the Presence of Firm Strategic Behavior," *Journal of Marketing Research*, 46 (2), 207-221.

Eraker, B. (2004), "Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices," *The Journal of Finance*, 59 (3), 1367-403.

Eraker, B., M. Johannes, and N. Polson (2003), "The Impact of Jumps in Volatility and Returns," *The Journal of Finance* 58 (3), 1269-300.

Fama, E. F., and J. D. MacBeth (1973), "Risk, Return, and Equilibrium: Empirical Tests," *Journal of Political Economy* 71, 607-636

Gamerman, Dani and Hedibert Freitas Lopes (2006), *Markov Chain Monte Carlo: Stochastic simulation for Bayesian Inference*, 2nd Edition, Chapman & Hall/CRC.

Gelman, Andrew, John B. Carlin, Hal S. Stern, Donald B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall/CRC Press.

Geman, S. and D. Geman (1984), "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6 (6), 721-41.

Gilks, Walter R., Sylvia Richardson and D. J. Spiegelhalter (2000), *Markov Chain Monte Carlo in Practice*, Chapman & Hall/CRC.

Greyserman, A., D. H. Jones, and W. E. Strawderman (2006), "Portfolio selection using hierarchical Bayesian analysis and MCMC methods," *Journal of Banking and Finance* 30 (2), 669-78.

Greene, William H. (2008), *Econometric Analysis*, 6th Edition, Pearson Prentice Hall.

Guadagni, P. M., J. D. C. Little (1983), "A logit model of brand choice calibrated on scanner data," *Marketing Science*, 2(3) 203 – 238.

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika* 57 (1), 97-109.

Jacquier, E., N. G. Polson, and P. E. Rossi (1994), "Bayesian Analysis of Stochastic Volatility Models," *Journal of Business and Economic Statistics* 12 (4), 371-89.

Jacquier, E. and N. G. Polson (2010), "Bayesian econometrics in finance", *The Oxford Handbook of Bayesian Econometrics*, edited by John Geweke, Gary Koop and Herman van Dijk, Oxford University Press.

Jacquier, E., N. G. Polson, and P. E. Rossi (2004), "Bayesian Analysis of Stochastic Volatility Models with Fat-Tails and Correlated Errors," *Journal of Econometrics* 122, 185-212.

Kim, S., N. Shephard, and S. Chib (1998), "Stochastic Volatility: Likelihodd Inference and Comparison with ARCH Models," *Review of Economic Studies* 65, 361-393.

Kim, J., G. M. Allenby and P. E. Rossi (2007), "Product Attributes and Models of Multiple Discreteness," *Journal of Econometrics* 138, 208-230.

Korteweg, A. (2010), "The Net Benefits to Leverage," *Journal of Finance* 65, 2137-2170.

Korteweg, A. (2011), "Markov Chain Monte Carlo Methods in Corporate Finance", Working paper.

Korteweg, A., and M. Sorensen (2010), "Risk and Return Characteristics of Venture Capital-Backed Entrepreneurial Companies," *Review of Financial Studies* 23(10), 3738-72.

Laplace, Pierre-Simon (1774). "Memoir on the probability of causes of events," *Mémoires de Mathématique et de Physique*, Tome Sixième. (English translation by S. M. Stigler 1986. *Statistical Science* 1(19):364–378).

Law, Averill M. and W. David Kelton (2000), *Simulation Modeling and Analysis*, 3rd Edition, McGraw-Hill.

Lewellen, J., S. Nagel, and J. Shanken (2010), "A Skeptical Appraisal of Asset-Pricing Tests," *Journal of Financial Economics* 96, 175-194.

Lo, A.W. and A.C. MacKinlay (1990), "Data-snooping Biases in Tests of Financial Asset Pricing Models," *Reviews of Financial Studies* 3, 431-468.

Markowitz, H. M. (1999), "Portfolio Selection," *Journal of Finance* 7, 77-91.

Montgomery, A. L. and P. Rossi (1999), "Estimating Price Elasticities with Theory-Based Priors," *Journal of Marketing Research* 36 (4), 413-423.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics* 21 (6), 1087-92.

Müller, P. and A. Pole (1998), "Monte Carlo Posterior Integration in GARCH Models," *Sankhya* 60, 127-44.

Nakatsuma, T. (2000), "Bayesian Analysis of ARMA-GARCH Models: A Markov Chain Sampling Approach," *Journal of Econometrics* 95, 57-69.

Robert, Christian P. and George Casella (1999), *Monte Carlo Statistical Methods*, Springer-Verlag New York.

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

Shin, S., S. Misra and D. Horsky, "Disentangling Preferences and Learning in Brand Choice Models," *Marketing Science*,31 (1), 115-137.

Tanner, M. A. and W. H. Wong (1987), "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association* 82 (398), 528-40.

Vrontos, I.D., P. Dellaportas, and D.N. Politis (2000), "Full Bayesian Inference for GARCH and EGARCH Models," *Journal of Business & Economic Statistics* 18 (2), 187-198.

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

Young, M. R. and P. Lenk (1998), "Hierarchical Bayes Methods for Multifactor Model Estimation and Portfolio Selection," *Management Science*, 44 (11) part 2, S111-S124.