Markov Chain Monte Carlo Methods in Corporate Finance

Arthur Korteweg *

November 25, 2011

Abstract

This chapter introduces Markov Chain Monte Carlo (MCMC) methods for empirical corporate finance. These methods are very useful for researchers interested in capital structure, investment policy, financial intermediation, corporate governance, structural models of the firm, and other areas of corporate finance. In particular, MCMC can be used to estimate models that are difficult to tackle with standard tools such as OLS, Instrumental Variables regressions and Maximum Likelihood. Starting from simple examples, this chapter exploits the modularity of MCMC to build sophisticated discrete choice, self-selection, panel data and structural models that can be applied to a variety of topics. Emphasis is placed on cases for which estimation by MCMC has distinct benefits compared to the standard methods in the field. I conclude with a list of suggested applications. Matlab code for the examples in this chapter is available on the author's personal homepage.

*Stanford University Graduate School of Business. I thank the editors, and Dirk Jenter, Kai Li, Michael Roberts, Morten Sørensen and Toni Whited for helpful comments and suggestions. All errors are my own. Matlab code to the examples in this chapter is available on the author's homepage at https://faculty-gsb.stanford.edu/korteweg/pages/DataCode.html

In the last two decades the field of empirical corporate finance has made great strides in employing sophisticated statistical tools to achieve identification, such as instrumental variables, propensity scoring and regression discontinuity methods. The application of Bayesian econometrics and in particular Markov Chain Monte Carlo (MCMC) methods, however, has been lagging other fields of finance such as fixed income and asset pricing, as well as other areas of scientific inquiry such as marketing, biology, and statistics. This chapter explores some of the many potential applications of this powerful methodology to important research questions in corporate finance.

With the current trend in the corporate finance literature towards more complex empirical models, MCMC methods provide a viable and attractive means of estimating and evaluating models for which classical methods such as least-squares regressions, GMM, Maximum Likelihood and their simulated counterparts are too cumbersome or computationally demanding to apply. In particular, MCMC is very useful for estimating non-linear models with high-dimensional integrals in the likelihood (such as models with many latent variables), or a hierarchical structure. This includes, but is not limited to, discrete-choice, matching and other self-selection models, duration, panel data and structural models, encompassing a large collection of topics in corporate finance such as capital structure and security issuance, financial intermediation, corporate governance, bankruptcy, and structural models of the firm. The MCMC approach thus opens the door to estimating more realistic and insightful models to address questions that have thus far been out of reach of empirical corporate finance.

To illustrate the method, I consider the effect of firm attrition on the coefficient estimates in typical capital structure panel data regressions. Firms disappear from the sample for non-random reasons such as through bankruptcy, mergers or acquisitions, and controlling for this non-random selection problem alters the estimated coefficients dramatically. For example, the coefficient on profitability changes by about 25%, and the coefficient on asset tangibility drops roughly in half. Whereas estimating this selection correction model is difficult with classical methods, the MCMC estimation is not particularly complex, requiring no more than standard probability distributions and standard regressions. I provide the Matlab code for this model on my personal website.[1]

The goals of this chapter are twofold. First, I want to introduce MCMC methods and provide a hands-on guide to writing algorithms. The second goal is to illustrate some of the many applications of MCMC in corporate finance. However, these goals come with a good deal of tension. Most sections in this chapter start with developing MCMC estimators for simple problems that have standard frequentist solutions that come pre-packaged in most

[1] https://faculty-gsb.stanford.edu/korteweg/pages/DataCode.html

popular software packages. The reason I discuss these examples is not because I think that researchers should spend their time coding up and running their own MCMC versions of these standard estimators, but rather to illustrate certain core principles and ideas. These simple examples then function as a stepping stone to more complex problems for which MCMC has a distinct advantage over the standard approaches such as least-squares regressions, GMM, and Maximum Likelihood (or where such approaches are simply not feasible). For example, section 3 starts with a standard probit model, focusing on the core concept of data augmentation. The modularity of MCMC allows us to extend this model to build a dynamic selection model in section 3.3 that is nearly impossible to estimate by Maximum Likelihood or other classical methods.

To aid readers interested in applying MCMC methods, I have provided Matlab code for all the numbered algorithms in this chapter on my personal webpage.[2] Apart from educational purposes, these examples can also be used as building blocks to estimate more complex models, thanks to the inherent modularity of MCMC. It is, for example, quite straightforward to add a missing data feature to a model by adding one or two steps to the algorithm, without the need to rewrite the entire estimation code.

At the core of MCMC lies the Hammersley-Clifford theorem, by which one can break up a complex estimation problem into bite-size pieces that usually require no more than standard regressions tools and sampling from simple distributions. Moreover, MCMC methods do not rely on asymptotic results but instead provides exact small-sample inference of parameters (and non-linear functions of parameters), and do not require optimization algorithms that often make Maximum Likelihood and GMM cumbersome to use.

The chapter is organized by modeling approach. Section 1 introduces Bayesian inference and MCMC estimation through a simple regression example. Section 2 introduces the concept of data augmentation through a missing data problem. Section 3 discusses limited dependent variable and sample selection models, currently the most widely used application of MCMC in corporate finance. Section 4 addresses panel data models and introduces the powerful tool of hierarchical modeling, and presents the application to capital structure regressions with attrition. Section 5 describes the estimation of structural models by MCMC, and in particular the concepts of Metropolis-Hastings sampling and Forward Filtering and Backward Sampling. Section 6 suggests a number of further applications in corporate finance for which MCMC is preferable to classical methods. Section 7 concludes.

[2]Other popular packages for running MCMC estimations are R, WinBugs and Octave (all can be downloaded free of charge). Many code examples for these packages can be found online.

1 Regression

1.1 A primer on Bayesian Inference

The Bayesian approach to statistical inference is somewhat different from the classical, frequentist setting. The fundamental difference is that the former approach views parameters as random variables whereas the latter views parameters as fixed characteristics of the population. Instead of finding the "best" estimate of the population parameters, Bayesians estimate the *posterior distribution* of the parameters. The posterior distribution is conditioned on all available data, and captures all that is known about the parameters based on the observed sample.

To fix ideas, consider the standard linear model

$$y = X\beta + \varepsilon,$$

where y is an $N \times 1$ vector of dependent variables, X is an $N \times k$ matrix of predictor variables, and the error term is distributed iid Normal, $\varepsilon \sim \mathcal{N}\left(0, \sigma^2 I_N\right)$. The parameters of interest are the vector of coefficients β and the error variance σ^2. The *joint* posterior distribution of the parameters given the observed data is written as $p(\beta, \sigma^2|X, y)$, and the *marginal* posterior distributions are $p(\beta|X, y)$ and $p(\sigma^2|X, y)$. Focusing on the marginal posterior distribution of β and applying Bayes' law, we may write

$$p\left(\beta|X, y\right) = \frac{p(y|X, \beta) \cdot p(\beta)}{p(y|X)}.$$

The denominator on the right hand side is merely a scaling constant since it does not depend on the model parameters. The information about the parameters is thus contained in the numerator, which is written as the product of the likelihood, $p(y|X, \beta)$ and the prior distribution, $p(\beta)$. The expression for σ^2 is analogous.

The posterior tells us everything there is to know about β based on the observed data. A popular point estimate to report is the mean of the posterior distribution (the "posterior mean"), which is the estimate that minimizes the mean squared error, $MSE = E\left[\left(\hat{\beta} - \beta\right)^2\right]$. Alternatively, the posterior median is optimal under the absolute loss function, $E\left[\left|\hat{\beta} - \beta\right|\right]$.

A statistic of particular interest is the posterior mode, sometimes referred to as the Maximum A Posteriori (MAP) estimate. With a flat (i.e. constant) prior, the posterior mode is exactly the Maximum Likelihood estimator.[3] Although the MAP underscores

[3]There is a technicality with proper versus improper priors. A proper prior is one that integrates to a finite number. The equivalence of the MAP to MLE holds for a flat but improper prior (i.e. not a

the link between Bayesian and classical Maximum Likelihood estimates, it is not typically reported because posterior distributions may be multi-modal. Moreover, the MAP is not invariant to reparametrization of the model.

When it comes to hypothesis testing, the Bayesian approach does not formally accept the classical concepts of significance and p-values. Instead, credible intervals of the posterior distribution can be roughly thought of as their Bayesian counterpart. For example, the (5,95%) credible intervals is the region between the 5th and 95th percentile of the posterior distribution.[4] If $\beta = 0$ is outside of the credible interval then there is not much support in the data for this particular parameter value.[5]

The prior distribution captures the researcher's information about the parameters prior to conducting the study. This prior information can be drawn from a variety of sources, for example, parameter restrictions from an economic model, estimates from previous studies, or plain economic intuition. The use of priors is often criticized for being the researcher's subjective choice. Naturally, many subjective choices are made in statistical inference, even in a frequentist setting, for example the choice of research design, model, variables, and the nature of the error terms, just to name a few. The prior is another such choice and, as in any decision, it is good practice to check the sensitivity of the results to the prior distribution. A necessary condition for a good prior is that it does not exert undue influence on the posterior distribution, but instead lets the data speak. This usually translates to having a prior that does not put a great deal of probability weight on a small subset of the parameter space. Such a prior is often called an "uninformative" prior, although this is a bit of a misnomer because any prior expresses *some* information about the parameters. For more discussion on priors, most Bayesian textbooks (e.g. [55, 89]), have a thorough description.

In the linear model, a typical choice of prior distributions is the Normal-Inverse Gamma prior:

$$\begin{aligned} \sigma^2 &\sim \mathcal{IG}(a, b), \\ \beta|\sigma^2 &\sim \mathcal{N}\left(\mu, \sigma^2 \cdot A^{-1}\right). \end{aligned}$$

Uniform distribution as it does not cover the entire parameter space). Improper priors can be problematic in MCMC algorithms, and in this chapter I will only use proper priors. For a detailed treatise on proper versus improper priors in MCMC, I refer the reader to Robert and Casella [88], p. 403-7. Under mild regularity conditions, Bayesian estimators are consistent and asymptotically Normally distributed.

[4]There are alternative ways of constructing credible intervals. For example, one could construct a 90% credible interval as the tightest interval that corresponds to 90% posterior probability, or a symmetric interval around the posterior mean.

[5]Berger and Delampady [4] argue that using credible intervals for hypothesis testing can be misleading (p.328). Instead, the proper decision-theoretic approach to hypothesis testing is to use Bayes Factors (e.g. [56]), which is a form of likelihood ratio test. I will not discuss the issue of hypothesis testing any further, but instead refer the reader to any thorough textbook on Bayesian estimation, such as Johannes and Polson [55] or Rossi et al. [89].

To gain some intuition on these priors, the Inverse Gamma (IG) distribution has non-negative support and a shape that looks roughly like the familiar F-distribution. The parameter a can be thought of as the number of observations, and b is roughly the sum of squared errors (SSE). The mean of the distribution is $b/(a-1)$, or (approximately) the SSE divided by the number of observations - similar to the standard calculation of variance. The variance of the IG distribution is $b^2/\left[(a-1)^2(a-2)\right]$. The first two moments are finite as long as $a > 2$. Unless otherwise noted, in this chapter I will set $a = 2.1$ and $b = 1$, corresponding to a mean of σ^2 of 0.91 and a variance of 8.26. In the prior for β, the $K \times K$ matrix A functions like a prior on $X'X$ and $A\mu$ can be thought of as a prior on $X'y$. Unless otherwise noted, I set $\mu = 0$ and $A = I_K \cdot 1/10,000$, such that the prior standard deviation on β is one hundred times σ^2.

With the Normal-Inverse Gamma priors, the posterior distribution turns out to have an analytical solution:

$$\begin{aligned} \sigma^2|X,y &\sim \mathcal{IG}(a+N, b+S) \\ \beta|\sigma^2,X,y &\sim \mathcal{N}\left(m, \sigma^2 \cdot (X'X+A)^{-1}\right) \end{aligned}$$

where $m = (X'X+A)^{-1}(X'y+A\mu)$ and $S = (y-Xm)'(y-Xm) + (m-\mu)'A(m-\mu)$. Note that the posterior distributions are of the same family as the priors, a phenomenon that is known as *conjugate* priors. After integrating out σ^2, the *marginal* posterior distribution of β conditional only on the observed data,

$$p(\beta|X,y) = \int p(\beta|\sigma^2,y,X)p(\sigma^2|y,X)d\sigma^2,$$

turns out to be the familiar t-distribution, which has similar contours as the conditional Normal distribution but with fatter tails.

There is an obvious parallel to OLS regression here: when $A \to 0$ (i.e. when the prior becomes *diffuse*) or as the number of observations $N \to \infty$, the posterior mean of β (i.e. m) converges to the OLS estimator $\hat{\beta} = (X'X)^{-1}X'y$. Moreover, S turns out to be the sum of squared errors from standard OLS regression. The Bayesian posterior mean estimators of β and σ^2 thus converge to the standard Maximum Likelihood estimators.[6]

1.2 Regression by Markov Chain Monte Carlo

An alternative way of learning about the marginal posterior distribution $p(\beta|X,y)$ is by drawing a sample from it through Monte Carlo simulation. Algorithm 1 explains the

[6]There is another interesting parallel: If we set the priors to $\mu = 0$ and $A = cI_K$ for a given constant c, then the posterior mean of β is the standard ridge regression estimator.

Algorithm 1 Regression

1. Draw $\sigma^2 \sim \mathcal{IG}(a+N, b+S)$
2. Draw $\beta|\sigma^2 \sim \mathcal{N}\left(m, \sigma^2 \cdot (X'X+A)^{-1}\right)$
3. Go back to step 1, repeat.

steps in detail (for ease of notation I do not write the conditioning on the observed data X and y in the algorithm). First, draw a realization of σ^2 from its posterior Inverse Gamma distribution.[7] Next, draw a realization of β using the draw of σ^2 from the first step, i.e. draw from the posterior distribution of $\beta|\sigma^2, X, y$. These two draws together form one draw of the *joint* posterior distribution $p(\beta, \sigma^2|X, y)$. Repeating these two steps many times then results in a sequence of draws from the joint posterior distribution. This sequence forms a Markov Chain (in fact the chain is independent here), hence the name Markov Chain Monte Carlo [32, 34]. This particular MCMC algorithm in which we can draw from exact distributions is also known as *Gibbs sampling*.

To illustrate the algorithm, I simulate a simple linear model with $N = 100$ observations and one independent variable, X, drawn from a standard Normal distribution, and no intercept. I set the true $\beta = 1$ and $\sigma = 0.25$, and use the priors as specified in the previous section. Using Matlab it takes about 3.5 seconds on a standard desktop PC to run 25,000 iterations of Algorithm 1. Figure 1 shows the histogram of the draws of β and σ. These histograms are the marginal posterior distributions of each parameter, numerically integrating out the other. In other words, the histogram of the 25,000 draws of β on the left-hand plot represents $p(\beta|X, y)$, and with enough draws converges to the analytical t-distribution. In the right-hand plot, the draws of σ^2 are first transformed to draws of σ by taking square roots before plotting the histogram. This highlights the ease of computing the posterior distribution of non-linear functions of parameters. This example may appear trivial here, but this transformation principle will prove very useful later.

The vertical lines in Figure 1 indicate the true parameter values. Note that the point estimates are close to the true parameters, despite the small sample and the *prior* means being centered far from the true parameter values. Moreover, the simulated posterior means coincide with the analytical Bayesian solutions. The (1,99%) credible intervals are $[0.910, 1.034]$ for β, and $[0.231, 0.290]$ for σ. The posterior standard deviation of β

[7] Drawing from an Inverse Gamma distribution in Matlab is fairly straightforward: simply draw X from a Gamma(a,1/b) distribution using the function `gaminv(rand, a, 1/b)`. The inverse 1/X then has an IG(a,b) distribution. See my online code examples for further details. One word of caution: the definition of b is unfortunately not standardized. In my notation I define the pdf of the Gamma(a,b) distribution as $f(x; a, b) = x^{a-1} \cdot e^{-x/b} / [b^a \Gamma(a)]$.

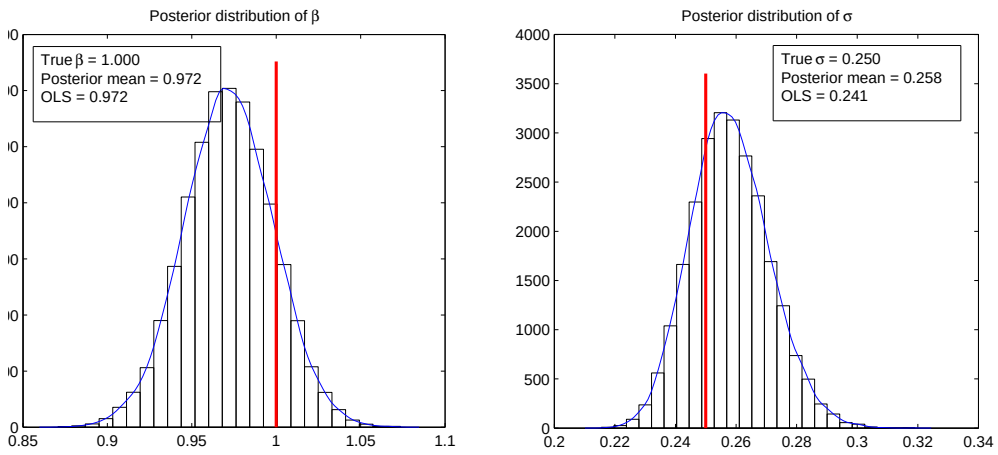

Figure 1: **Posterior distribution of regression parameters.**
Histograms of the 25,000 parameter draws of the standard regression model estimated by MCMC on a simulated dataset of 100 observations. The vertical lines indicate the true parameter values that were used to generate the data.

is 0.026, compared to a standard error of 0.025 from OLS regression. The difference is due to the MCMC estimates being small-sample estimates that do not rely on asymptotic approximations, unlike standard errors. After all, the very definition of the posterior distribution implies that the estimates are conditioned on the observed data only, not an imaginary infinitely large dataset. This allows for exact inference, which may be quite different from the asymptotic inference of classical methods, especially in smaller datasets.

2 Missing data

To make the inference problem more interesting, suppose that some of the observations in y are missing at random (I postpone the problem of non-randomly missing data until the next section). The problem of missing data is widespread in corporate finance, even for key variables such as investments and leverage. For example, for the 392,469 firm-year observations in Compustat between 1950 and 2010, capital expenditures is missing for 13.9% of the observations. Debt issuance has been collected since 1971 but 14.1% of the 348,228 firm-year observations are missing, whereas market leverage (book debt divided by book debt plus market value of equity) is missing 22.4% of the time. For R&D expenditures the missing rate is over 50%. Even a canonical scaling variable such as total assets is missing around 5% of the time. As I will illustrate below, MCMC provides a convenient way of dealing with the missing data problem.

With missing data one loses the Bayesian analytical solution to the posterior distribution. However, the sampling algorithm from the previous section needs only one, relatively

Algorithm 2 Missing data

1. Draw the missing y_i^* for all i with missing y_i, treating β and σ^2 as known:

$$y_i^* | \beta, \sigma^2 \sim \mathcal{N}\left(X'\beta, \sigma^2\right)$$

2. Draw β, σ^2 from a Bayesian regression with Normal-IG priors, treating the y^*as observed data. The posterior distributions are:

$$\begin{aligned} \sigma^2 | y^* &\sim \mathcal{IG}(a+N, b+S) \\ \beta | \sigma^2, y^* &\sim \mathcal{N}\left(m, \sigma^2 \cdot (X'X+A)^{-1}\right) \end{aligned}$$

3. Go back to step 1, repeat.

minor, modification based on the important concept of data augmentation [96] in order to deal with this issue. Think of the missing observations, denoted by y^*, as parameters to be estimated along with the regression parameters. In other words, we augment the parameter vector with the latent y^*, and sample from the joint posterior distribution $p\left(\beta, \sigma^2, y^* | X, y\right)$.

The key to sampling from this augmented posterior distribution is the Hammersley-Clifford theorem. For our purposes, this theorem implies that the complete set of conditional distributions $p\left(\beta, \sigma^2 | y^*, X, y\right)$ and $p\left(y^* | \beta, \sigma^2, X, y\right)$ completely characterizes the joint distribution. Algorithm 2 shows that, unlike the joint distribution, the complete conditionals are very straightforward to sample from: $p\left(y^* | \beta, \sigma^2, X, y\right)$ is a Normal distribution (and each missing observation can be sampled independently since the error terms are iid), and $p\left(\beta, \sigma^2 | y^*, X, y\right)$ is simply Algorithm 1, a Bayesian regression treating the missing y^* as observed data. This gives a first taste of the modularity of the MCMC approach: we go from a standard regression model to a regression with missing data by adding an extra step to the algorithm. Note again that I suppress the conditioning on the observed data in the algorithm.

Denoting by $\left\{\sigma^2\right\}^{(g)}$ the draw of σ^2 in cycle g of the MCMC algorithm, the MCMC algorithm thus starts from initial values $\{\beta\}^{(0)}$ and $\left\{\sigma^2\right\}^{(0)}$, and cycles between drawing y^*, σ^2 and β, conditioning on the latest draw of the other parameters:

$$\{\beta\}^{(0)}, \left\{\sigma^2\right\}^{(0)} \rightarrow \{y^*\}^{(1)} \rightarrow \left\{\sigma^2\right\}^{(1)} \rightarrow \{\beta\}^{(1)} \rightarrow \{y^*\}^{(2)} \rightarrow \ldots$$

The resulting sequence of draws is a Markov Chain with the attractive property that, under mild conditions, it converges to a stationary distribution that is exactly the augmented joint posterior distribution. This is the essence of MCMC. Figure 2 shows the first 50 draws of Algorithm 2, using the same regression model as the previous section (with true $\beta = 1$

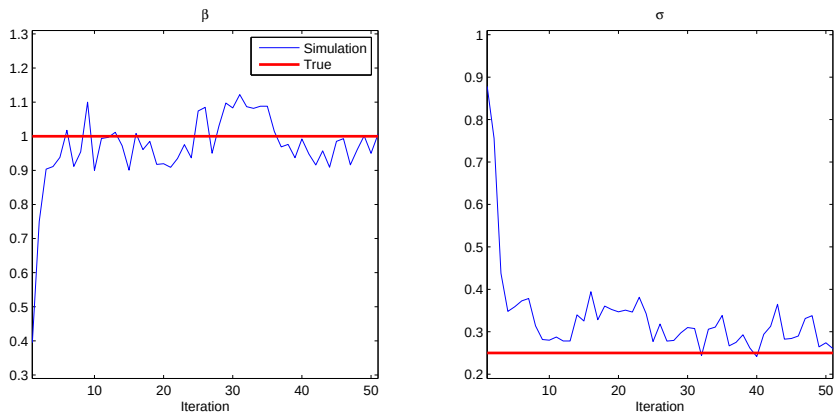

Figure 2: **Convergence of the MCMC chain.**
Plot of the first 50 iterations of the Markov Chain of the estimation of the missing data regression model in Algorithm 2, estimated on a simulated dataset of 100 observations, of which 50 are dropped at random. The horizontal lines indicate the true parameter values that were used to generate the data.

and $\sigma^2 = 0.25$), but randomly dropping half of the observations of y. The convergence to the stationary distribution is most noticeable for σ^2, and quite rapid for this particular model. The period of convergence is called the "burn-in" period and is dropped before calculating parameter estimates and other properties of the posterior distribution.

For many problems the likelihood function is not globally concave and has multiple local maxima. For such problems the chain needs to run for a larger number of cycles in order to fully explore the posterior distribution. As a general rule, the MCMC algorithm is more hands-off than Maximum Likelihood, which requires a great deal of manual work by the researcher to make sure that a global optimum is reached, for example through the use of different starting values, applying a variety of maximization algorithms, or simulated annealing routines. In rare occasions, it is possible even for the MCMC chain to get "stuck" in a local maximum, so it is still good practice to try different starting values. This is also helpful in determining when the chain has converged (Gelman and Rubin [33] develop a convergence test based on the within and between variance of multiple MCMC chains), and does not waste much computation time because the post-convergence draws of the different chains can be combined in order to obtain a better approximation to the posterior distribution.

Table 1 shows parameter estimates from two OLS regression approaches to the missing data problem, as well as MCMC estimates. The first OLS results drop the observations for which y is unobserved altogether. The second set of OLS estimates fill in the unobserved y^* using the point estimates of β from the dropped observations. In other words, $y_i^* = x_i'\hat{\beta}$. Unlike other common fill-in schemes such as using the sample average, this results in unbiased estimates of β.

	True	OLS		MCMC	
		Drop	Filled	Drop	Alg 2
β	1	1.012	1.012	1.013	1.011
		(0.047)	(0.025)	(0.058)	(0.057)
σ	0.25	0.239	0.171	0.298	0.297
		-	-	(0.029)	(0.032)

Table 1: **Missing dependent variable regressions.**
Estimates of a missing data regression model based on a simulated dataset of 100 observations, randomly dropping 50% of the dependent variable data (but not the regressor). The true coefficients are shown in the first column (there is no intercept). The OLS columns estimate the model by dropping the missing observations altogether (the column labeled "Drop"), and filling in the missing data using fitted values from the "Drop" regression (The column labeled "Filled"). The MCMC estimates in the "Drop" column uses the standard Bayesian regression of Algorithm 1, dropping the missing observations. The final column uses Algorithm 2, which simulates and integrates out the missing observations. The MCMC estimates use 1,000 burn-in cycles followed by 25,000 cycles to sample the posterior distribution. Standard errors for the OLS estimates and posterior standard deviations for the MCMC estimates are in brackets.

The key issue here is one of statistical efficiency: dropping the unobserved data altogether ignores the information in X that is contained in the dropped observations, while filling in the missing data with point estimates understates standard errors by ignoring the prediction variance in the filled-in data. The latter problem is evident from the fact that the standard errors as well as the estimates of the residual standard deviations of the filled-in OLS regressions are considerably lower compared to the OLS regressions that drop the data with missing observations. The MCMC algorithm solves both these problems by using all observations on X while accounting for prediction variance by filling in different values for y^* every time we take a draw of β and σ^2. For comparison, the first MCMC column shows the MCMC version of the regression dropping the observations with missing y's. The posterior standard deviations are larger than the OLS standard errors because they are small-sample rather than asymptotic estimates. The last column shows the results from Algorithm 2, accounting for the information in the dropped X while also accounting for the uncertainty of the missing y^*.

It is evident from Table 1 that integrating out the missing y^* is not particularly helpful in this example. However, Table 2 shows that the benefits are more substantial when allowing for missing observations on the *explanatory* variables in a multiple regression. Simulating and integrating out the missing variables leads to parameter estimates that are generally closer to the true parameters than the other methods, even for a relatively low correlation between the explanatory variables of 0.15. Moreover, since the full information about the dependent and non-missing explanatory variables is exploited, the estimates have lower posterior standard deviations, i.e. they are more precise, compared to the MCMC estimates that drop the observations with some missing data altogether. The algorithm for tackling this problem is very similar to Algorithm 2, essentially simulating

	True	OLS	GLS	MCMC	
		Drop	Filled	Drop	Alg 2
β_1	0.5	0.493	0.416	0.493	0.476
		(0.038)	(0.002)	(0.042)	(0.039)
β_2	-0.5	-0.425	-0.426	-0.424	-0.474
		(0.044)	(0.003)	(0.049)	(0.039)
σ	0.25	0.251	0.262	0.280	0.282
		-	-	(0.020)	(0.025)

Table 2: **Missing explanatory variable regressions.**
Estimates of a missing data regression model based on a simulated dataset of 100 observations with two explanatory variables, $y = x_1\beta_1 + x_2\beta_2 + \varepsilon$, and randomly dropping 50% of the observations on x_2. The true coefficients are shown in the first column. The explanatory variables have a correlation coefficient of 0.15. The first column estimates the model by dropping the missing observations altogether, and the second column uses the Griliches [38] GLS method to fill in the missing data, using a regression of the variable with missing data on the other explanatory variable. The MCMC estimates in the "Drop" column use the standard Bayesian regression of Algorithm 1, dropping the missing observations, whereas the final column uses a version of Algorithm 2 to simulate and integrate out the missing observations. The MCMC estimates use 1,000 burn-in cycles followed by 25,000 cycles to sample the posterior distribution. Standard errors for the OLS estimates and posterior standard deviations for the MCMC estimates are in brackets.

the missing explanatory variables from a regression of the variable with missing data on the other variables. An example of a corporate finance application to such a problem is in Frank and Goyal [30], who impute missing factors in leverage regressions using an MCMC algorithm.

Other, more complex, cases also promise better results, for example if y follows a time-series process but has missing data gaps in the series. An illustration of this kind is found in Korteweg [58], who uses panel data for corporate bond and equity values to estimate the net benefits to leverage, where a non-trivial fraction of the corporate bond values are unobserved. Another avenue for improving performance is to sample the empirical distribution of the residuals, instead of imposing the Normal distribution, to obtain results that are more robust to distributional assumptions. For further information on using Bayesian methods for missing data, see Rubin [90] and Graham and Hirano [37].

3 Limited Dependent Variable and Selection models

In the previous section I assumed that data is missing at random. If data is instead missing for a reason, it becomes necessary to specify the model by which observations are selected in order to obtain estimates of the parameters of interest. More generally, selection models are useful for addressing many questions in corporate finance such as the effect of underwriter choice on bond issue yields, the diversification discount due to conglomeration choice, and the impact of bank versus pubic debt financing on cost of

capital (see Li and Prabhala [70] for a comprehensive overview and references).

I start this section with a simple probit example. The probit model serves as a basis to developing an estimation algorithm for the Heckman selection model. Since both probit and Heckman have canned Maximum Likelihood-based modules in popular statistics packages such as Stata, this may not sound very exciting. However, in section 3.3 and beyond, I introduce extensions to dynamic selection, matching models and switching regressions, that are very difficult to estimate with Maximum Likelihood, but are quite feasible with MCMC both from an ease of implementation as well as a computational perspective.

3.1 Probit model

In the standard probit model, y_i has two possible outcomes, zero and one. The probability of observing $y_i = 1$ is:

$$pr(y_i = 1) = \Phi(x_i\beta),$$

where $\Phi(\cdot)$ is the standard Normal cdf. Observations are assumed to be iid. Probit models have been used in corporate finance to estimate, for example, the probability of issuing debt or equity [50], takeovers [5], bankruptcy [86], and the firing of ceo's [54].

The estimation goal is to find the posterior distribution of β given y and X. It will prove convenient to rewrite the model in the following way:

$$\begin{aligned} y_i &= \mathbb{I}_{\{w_i \geq 0\}} \\ w_i &= x_i\beta + \eta_i \end{aligned}$$

with $\eta_i \sim \mathcal{N}(0,1)$, iid. The auxiliary selection variable w is unobserved. If $w \geq 0$ then y equals one, otherwise y equals zero. Augmenting the posterior distribution with w, MCMC Algorithm 3 (from Albert and Chib [3]) shows how to sample from the joint posterior distribution of β and w, conditional on the observed data. The algorithm cycles between drawing from the complete conditionals $w|\beta$ and $\beta|w$, which by the Hammersley-Clifford theorem fully characterize the joint posterior distribution.

In step 1, when y equals one, w must be greater than zero, and the distribution of w is therefore truncated from below at zero. I denote this lower-truncated Normal distribution by $\mathcal{LTN}$. Similarly, $\mathcal{UTN}$ is the upper truncated Normal distribution, again truncated at zero. Step 2 draws from the posterior distribution of coefficients in a Bayesian regression, as in Algorithm 1 but fixing the variance of the error term to unity.

An advantage of MCMC for the probit model is in the calculation of non-linear functions of parameters. Researchers are often interested in the partial effects, $\partial pr(y =$

Algorithm 3 Probit model

1. Draw $w_i|\beta$ for all i:

   (a) for $y_i = 1$:

   $$w_i|\beta \sim \mathcal{LTN}(x_i\beta, 1)$$

   (b) for $y_i = 0$:

   $$w_i|\beta \sim \mathcal{UTN}(x_i\beta, 1)$$

2. Draw $\beta|w$ from a Bayesian regression of w on X, with Normal priors on β and known variance = 1:

   $$\beta|w \sim \mathcal{N}\left((X'X + A)^{-1}(X'w + A\mu), (X'X + A)^{-1}\right)$$

3. Go back to step 1, repeat.

$1|x)/\partial x = \phi(x'\beta)\beta$, which are highly non-linear functions of β. With the standard Maximum Likelihood approach the asymptotic distribution of the partial effects has to be approximated using the Delta method. With MCMC we obtain a sample from the exact posterior distribution without relying on asymptotics or approximations by simply calculating the partial effect for each draw of β (discarding the burn-in draws). We can then compute means, variances, intervals etc.

Many extensions of Algorithm 3 have been developed, for example to the multivariate probit model with several outcome variables [15], the multinomial probit that allows more than two discrete outcomes [75], and the multinomial-t and multinomial logit models [16]. In the next section I discuss another extension of the probit model, the classic Heckman selection model.

3.2 Heckman selection model

In the Heckman (also known as Tobit type 2) selection model, y_i is no longer a binary variable but can take on continuous outcomes:

$$\begin{aligned} y_i &= x_i\beta + \varepsilon_i \\ w_i &= z_i\gamma + \eta_i \end{aligned}$$

The outcome variable y_i is observed only when $w_i \geq 0$. The error terms are distributed iid jointly Normal with zero means, $var(\varepsilon_i) = \sigma^2$, $var(\eta_i) = 1$, and correlation ρ. The top equation is referred to as the outcome equation, and the bottom equation is the selection equation.

The Heckman selection model can be used to control for endogenously missing data and self-selection by firms. For example, the choice to issue equity may depend on the type of firm. If there is an unobserved component to firm type, then selection cannot be controlled for by covariates in the outcome equation alone. In panel data applications it is common practice to use fixed effects to control for selection, but these do not control for the fact that the reasons for selection can (and often do) change over time. Selection models do allow for that possibility.

To estimate the Heckman model by MCMC, we decompose ε_i into a component that loads on η_i and an orthogonal component ξ_i:

$$\begin{aligned} y_i &= x_i\beta + \delta \cdot \eta_i + \sigma_\xi \cdot \xi_i \\ w_i &= z_i\gamma + \eta_i \end{aligned}$$

where $\delta = \sigma \cdot \rho$ is the covariance between ε and η, and $\sigma_\xi = \sigma \cdot \sqrt{1-\rho^2}$ is the conditional standard deviation of $\varepsilon|\eta$. From this representation it follows immediately that the selection equation cannot be ignored if ρ (and hence δ)is not equal to zero. Consider the expected value of y_i if it is observed:

$$E\left[y_i|w_i \geq 0, data\right] = x_i\beta + \delta E\left[\eta_i|\eta_i \geq -z_i\gamma\right]$$

Ignoring the selection equation (dropping $\delta \cdot \eta$ in the observation equation) thus introduces an omitted variables bias if $\rho \neq 0$ [44]. The omitted variable, $E\left[\eta_i|\eta_i \geq -z_i\gamma\right] = \phi(\cdot)/\Phi(\cdot)$, is called the inverse Mills ratio.

MCMC Algorithm 4 (based on Li [67]) shows how to draw from the posterior distribution of the parameters augmented with the latent selection variable, w, and the missing observations, y^*. The algorithm essentially combines the randomly missing data routine with the probit estimation.

From the sampled parameters it is straightforward to use non-linear transformations to recover the posterior distribution of ρ and σ, as well as treatment effects, analogous to the calculation of partial effects in the probit model.

The typical approach to estimating Heckman models is the two-step estimator [44]: In the first stage we estimate the selection equation using a probit model, and in the second stage we plug the fitted inverse Mills ratio from the first stage into the outcome equation to correct for the omitted variable bias. This estimator generates inconsistent estimates for the covariance matrix in the second stage, and correct standard errors have to be computed from an asymptotic approximation or through a bootstrap. Full information estimators

Algorithm 4 Heckman selection model

1. Draw $w_i, y_i^* | \beta, \gamma, \delta, \sigma_\xi^2$

   (a) for y_i observed:

$$w_i | \beta, \gamma, \delta, \sigma_\xi^2 \sim \mathcal{LTN}\left(z_i\gamma + \rho \cdot \left[\frac{y_i - x_i\beta}{\sqrt{\delta^2 + \sigma_\xi^2}} \right], 1 - \rho^2 \right)$$

$$\text{where } \rho = \delta / \sqrt{\delta^2 + \sigma_\xi^2}$$

   (b) for y_i not observed:

$$\begin{aligned} w_i | \beta, \gamma, \delta, \sigma_\xi^2 &\sim \mathcal{UTN}\left(z_i\gamma, 1 \right) \\ y_i^* | w_i, \beta, \gamma, \delta, \sigma_\xi^2 &\sim \mathcal{N}\left(x_i\beta + \delta\left[w_i - z_i\gamma \right], \sigma_\xi^2 \right) \end{aligned}$$

2. Draw $\beta, \gamma | w, y^*, \delta, \sigma_\xi^2$ from a Bayesian Seemingly Unrelated Regression (Zellner, 197x) of $[y; w]$ on $[X; Z]$, with Normal priors on β and γ and known covariance matrix Ω:

$$\beta, \gamma | w \sim \mathcal{N}\left(\left(C'\Omega^{-1}C + A \right)^{-1} \left(C'\Omega^{-1}\left[y; w \right] + A\mu \right), \left(C'\Omega^{-1}C + A \right)^{-1} \right)$$

   where

$$\Omega = \begin{bmatrix} \sigma_\xi^2 + \delta^2 & \delta \\ \delta & 1 \end{bmatrix} \otimes I_N \qquad \text{and} \qquad C = \begin{bmatrix} X & 0 \\ 0 & Z \end{bmatrix}$$

3. Draw $\delta, \sigma_\xi^2 | \beta, \gamma, w, y^*$ from a Bayesian regression of $y - X\beta$ on $w - Z\gamma$, with Normal-IG priors (see Algorithm 1).

4. Go back to step 1, repeat.

(MCMC and the Maximum Likelihood estimator) exhibit better statistical properties but are often criticized for their sensitivity to the Normality assumption. Robust and non-parametric estimators have been proposed to deal with this issue (e.g. [46, 73, 74]), but tend to be limited in the types of models they can estimate. For example, Manski's [74] model is limited to two regressors. In contrast, the MCMC algorithm is more flexible. Van Hasselt [99] extends the algorithm to accomodate a mixture of Normals in the error terms without losing the generality of the model. Mixtures of Normals are able to generate many shapes of distributions such as skewness, kurtosis, and multimodality, and the algorithm lets the data tell you what the shape of the error distribution is. Van Hasselt also shows how to estimate the number of mixture components, something that is very difficult to do with frequentist methods.[8]

The Heckman selection algorithm outlined above can be adapted to estimate many related models. For example, Chib [14] develops an MCMC algorithm to estimate the Tobit censoring model where y takes on only non-negative values. Tobit censoring is relevant in corporate finance applications because many variables are naturally non-negative, such as gross equity issuance, cash balances and investment, and ignoring any censoring (for example from irreversible investment bounding capital expenditures at zero) may mask certain causal relationships of interest. Double-censoring can also be accommodated, which is of particular interest in corporate finance as, for example, it has recently been shown that accounting for the fact that leverage (measured gross of cash) is bounded between zero and one is important for statistical estimates of speed-of-adjustment models [11, 52]. Similarly, the outcome variable may be qualitative (zero or one) to model dichotomous outcomes such as default vs no default, or merger vs no merger.

The standard selection model has only two possible outcomes from selection: a data point is either observed or not observed. In many corporate finance applications there are multiple possible outcomes. For example, a company can choose not to raise debt financing, choose to issue public bonds, or to obtain a bank loan. Obrizan [83] shows how to estimate the selection model with a multinomial probit in the selection equation by MCMC. Classical estimation of this model is possible as well, but requires numerical integration of the likelihood using quadrature or simulation. I will return to this issue below. In the remainder of this section I introduce three other extensions to the Heckman model that have many potential applications in corporate finance but are very difficult to estimate with traditional methods.

[8] A common concern with the standard selection model is that it is identified from distributional assumptions only, unless one employs an instrument that exogenously changes the probability of being selected but is orthogonal to ε, the shock in the observation equation [46]. Relaxing the Normality assumption may loosen the distributional assumptions somewhat, but should not be seen as a substitute for an instrument.

3.3 Dynamic selection

The standard Heckman model assumes that the error terms in the selection equation are independent across units of observation. Under this assumption the unobserved data points are only informative about γ, the parameters of the selection equation. They carry no further information about the parameters of the outcome equation, β. In many corporate finance applications this is not the case. For example, a firm may be more inclined to issue equity because a peer firm is doing the same, or because some common unobserved factor induces both firms to issue. For the purpose of illustration, consider the case of two firms. The outcome of the first company is unobserved and the outcome of the second is observed. The expected value of the outcome of the second firm is

$$E\left[y_2|w_2 \geq 0, w_1 < 0, data\right] = x_2\beta + \delta E\left[\eta_2|\eta_2 \geq -z_2\gamma, \eta_1 < -z_1\gamma\right]$$

Unlike the standard Heckman model, if the $\eta's$ are correlated then firm 1 carries information about the conditional mean of y_2 despite firm 1's outcome being unobserved. In other words, the fact that firm 1 is unobserved matters for the conditional mean of firm 2. With two firms the expectation is a two-dimensional integral. With thousands of firms the integral becomes very high-dimensional, and with the current state of computing power it is too time-consuming to evaluate in a Maximum Likelihood estimation, or even to compute the inverse Mills ratio in the two-step procedure.[9]

A similar estimation issue arises when the outcome variable follows an autoregressive distributed lag (ADL) model. For example, consider the following ADL process for an individual firm:

$$y_t = \lambda y_{t-1} + x_t\beta + \varepsilon_t$$

Assume that the error terms are temporally independent (but ε and η are still contemporanesouly correlated so that $\delta \neq 0$). In a two-period setting, if the outcome at time 1 is unobserved but observed at time 2, the conditional mean of the outcome at time 2 is

$$E\left[y_2|w_2 \geq 0, w_1 < 0, data\right] = x_2\beta + \lambda E\left[y_1|w_2 \geq 0, w_1 < 0, data\right] + \delta E\left[\eta_2|\eta_2 \geq -z_2\gamma\right]$$

With $\lambda = 0$ this works out to the standard Heckman correction. With non-zero λ we get a similar integration issue as in the cross-sectional case, because the value of y_1 depends on the realized value of y_2, due to the ADL process. The resulting model is thus a dynamic generalization of the standard selection model. Autocorrelation in the error term results

[9]The usual way to numerically integrate out the latent variables in Maximum Likelihood is through quadrature methods, which are effective only when the dimension of the integral is small, preferably less than five[98]. The alternative, Simulated Maximum Likelihood (SML), is less computationally efficient than MCMC. I will discuss this in more detail in section 4.

in similar estimation problems, even without the lagged dependent variable.

Korteweg and Sørensen [61] tackle the estimation problem of the dynamic selection model using an MCMC algorithm, and apply it to estimate the risk and return to venture capital funded entrepreneural firms. The outcome variable is the natural logarithm of a start-up's market value, which follows a random walk (the ADL process above with $\lambda = 1$). However, the valuation is only observed when the company obtains a new round of funding. The probability of funding depends strongly on the firm's valuation, and this gives rise to the selection problem. Korteweg and Sørensen develop an MCMC algorithm to estimate the model in a computationally efficient way.[10] In a follow-up paper, Korteweg and Sørensen [62] apply the model to estimating loan-to-value ratios for single-family homes, as well as sales and foreclosure behavior and home price indices. They extend the model to include multiple selection equations, in order to capture the probability of a regular sale versus a foreclosure sale.

The dynamic selection model is very versatile, and can be applied to virtually any linear asset pricing model with endogenous trading. The model is also applicable to a variety of corporate finance problems since many of the standard variables follow ADL processes, and, in addition, autocorrelation in the error term is a common occurrence. Moreover, the estimation problem is not unique to the selection model, but generalizes to the Tobit model and other censoring problems. For example, investment follows an ADL but is censored at zero (if irreversible), leading to similar integration issues as the dynamic selection model. Other examples include cash balances, default or M&A intensity, and leverage, where the latter is subject double censoring as mentioned above.

3.4 Switching regressions

In a switching regression, y is always observed, but the parameters of the outcome equation depend on whether w is above or below zero:

$$y_i = \begin{cases} y_{i0} = x_{i0}\beta_0 + \varepsilon_{i0} & \text{if } w_i > 0 \\ y_{i1} = x_{i1}\beta_1 + \varepsilon_{i1} & \text{if } w_i \leq 0 \end{cases}$$

For example, Scruggs [91] considers the market reaction to calls of convertible bonds, which can be "naked" (without protection) or with the assistance of an underwriter that guarantees conversion at the end of the call period. We observe the announcement reaction, y, under either call method but the choice of method is endogenous through the correlations between η, ε_0 and ε_1. Unlike standard average announcement effect studies,

[10] A detailed description of the algorithm as well as Matlab and C++ code to implement it can be found on my personal webpage.

the β_0 and β_1 parameters in the switching regressions reflect announcement effects *conditional* on the endogenously chosen call method, as advocated by Acharya [1], Eckbo, Maksimovic, and Williams [26], Maddala [72], and Prabhala [85]. In other words, the parameters capture the counterfactual of what would have happened if a firm had chosen the unobserved alternative. As such, we can use the model to analyze treatment effects. As in the Heckman model, without an instrument the model is identified from parametric assumptions alone, and imposing an exclusion restriction is helpful to achieve non-parametric identification. Scruggs assumes that the error terms have a fat-tailed multivariate t-distribution and estimates the model using an MCMC algorithm.

Switching models have many potential applications in corporate finance. For example firm's cash and investment policies may be different in growth versus recession regimes, or hiring and firing intensities of CEO's may vary depending on the state of the firm.

It is possible to estimate a switching regression using classical methods [44, 63], but this is cumbersome at best. The MCMC approach is flexible and easy to extend to more complex models. For example, Li and McNally [69] use MCMC to estimate a switching regression with multiple outcome equations within each regime. They apply their model to the choice of share repurchase method, modeling the percentage of shares bought, the announcement effects and the tender premium (if the repurchase is by tender offer). This model can in turn be extended to outcome equations that are not all continuous, but instead may be composed of a mix of continuous, truncated and discrete outcomes. Another logical extension is to have more than two regimes since in many cases corporate managers face a decision between multiple options (similar to the multinomial selection model). Such models become increasingly intractable with classical methods, but are quite manageable using MCMC.

3.5 Matching models

A prevalent form of matching in corporate finance is the endogenous two-sided matching between two entities. For example, firms match with banks for their financing needs, and CEO's match with the firms they run. Sørensen [93] develops a model in which venture capitalists (VCs) match with entrepreneurs. He asks whether the better performance of experienced VCs is driven by sorting (more experienced VCs pick better firms) or influence (more experienced VCs add more value). Since some of the dimensions along which sorting occurs are unobserved, the resulting endogeneity problem makes identification more tricky. The economics of the problem makes finding an instrument very difficult, so Sørensen develops a structural model that exploits the fact that investors' decision to invest depends on the other agents in the market, whereas the outcome of the investment

does not. This provides the exogenous variation needed for identification.

The resulting model is prohibitively time-consuming to estimate by Maximum Likelihood because investment decisions interact. If one investor invests in a start-up, then other investors cannot. This implies that the error terms in the model are not independent and have to be integrated jointly in order to compute the likelihood function. Given that there are thousands of investments, such an extremely high-dimensional integral is computationally infeasible at present. Sørensen develops a feasible MCMC procedure to estimate the model, which is computationally much quicker then Maximum Likelihood.

Later studies [10, 84] use a similar MCMC methodology to study the matching of targets and acquirers in M&A, and the matching between banks and firms.

The next section on panel data dives deeper into the benefits of MCMC methods for formulating feasible estimators that perform high-dimensional integration in a computationally efficient way.

4 Panel data

In corporate finance one often observes the actions of a set of agents (companies, CEO's etc) over time. Such panel datasets are a rich source of identification, but also come with certain empirical challenges. The standard issues in classical estimation of panel data models are the assumptions regarding asymptotics (whether we assume that N or T approaches infinity) and the related incidental parameters problem [80][11], the initial values problem [45], and the Hurwicz asymptotic bias for ADL type models (also known as Nickell bias or, when applied to predictive regressions, Stambaugh bias). The Bayesian approach avoids many of these pitfalls. For example, asymptotic assumptions are unnecessary in the Bayesian paradigm since one conditions on the observed data only, and the initial values problem is easier to handle since we can treat it like a missing data problem. Moreover, MCMC methods allow the estimation of a wider variety of panel data models, as I will discuss below.

4.1 Random effects probit

Consider the panel data extension of the probit model with random effects (RE):

$$\begin{aligned} y_{it} &= \mathbb{I}_{\{w_{it} \geq 0\}} \\ w_{it} &= x_{it}\beta + \alpha_i + \eta_{it} \end{aligned}$$

[11] The incidental parameters problem in the panel data context states that individual fixed effects are not estimated consistently for fixed T, which results in inconsistent estimates of the parameters of interest. In some cases a transformation (such as first-differencing to cancel out the fixed effects) can resolve the problem, but these are rarely found outside of the linear and logit models.

For example, $pr(y_{it} = 1)$ could represent the probability that firm i goes bankrupt at time t. The unit-specific intercept, α_i, is assumed to be randomly generated from a Normal distribution with mean zero and variance τ^2, and is uncorrelated with $\eta_{it} \sim \mathcal{N}\left(0, 1-\tau^2\right)$. This "random effect" controls for time-invariant, unobserved heterogeneity across units.[12] The parameters, β, are therefore identified from the time-series variation within firms.[13]

It is useful to think of the structure of the panel probit model as a hierarchy, where each level builds upon the previous levels:

$$\begin{aligned} \tau^2 &\sim \mathcal{IG}(a, b) \\ \alpha_i|\tau^2 &\sim \mathcal{N}\left(0, I_N \cdot \left(1-\tau^2\right)/\tau^2\right) \\ w_{it}|\alpha_i, \tau^2 &\sim \mathcal{N}\left(x_{it}\beta + \alpha_i, 1-\tau^2\right) \end{aligned}$$

This hierarchy can be extended to as many levels as desired (e.g. industry-company-executive-year data). Hierarchical models [71] are useful in many corporate finance settings, and MCMC methods are very well suited for estimating these models, due to the complete conditional structure of the algorithm. By breaking up the problem in simple regression steps based on its hierarchical structure one can estimate models for which even the act of writing down the likelihood function becomes an arduous task. This allows one to compute correct standard errors and perform hypothesis testing without resorting to standard shortcuts such as two-stage estimators.

Algorithm 5 shows how to extend the probit Algorithm 3 to estimate the panel probit model. Step 1 and 2 follow straight from Algorithm 3. Step 3 jointly draws a set of α's by regressing $w_{it} - x_{it}\beta$ on a set of dummies, one for each firm (note that the prior means are zero). Step 4 estimates the variance of the α's, again in regression form. Note that the Algorithm follows the hierarchy of the model.

Besides the relative ease of programming[14] there is also a computational advantage to MCMC in models with high-dimensional integrals over many latent variables. To appreciate why non-linear panel data models such as the panel RE probit are difficult to

[12] Alternatively, one can think of the random effects as a form of error clustering. Note that In Stata the "cluster" command gives larger standard errors than the RE estimates, because Stata only considers the residual idiosyncractic error after removing the group error component.

[13] The random effects estimator is different from the fixed effects estimator, which is typically estimated using dummy variables. The random effects estimator dominates the fixed effects estimator in mean-squared error [27, 94], whereas the benefit of the fixed effects estimator is that it allows the unit-specific means to be correlated with the other explanatory variables. Mundlak [78] develops a correlated random effects model by specifying $\alpha_i = \bar{x_i}\gamma + u_i$, where $\bar{x_i}$ is the time-series average of x_{it}, and u_i is an othorgonal error term. Chamberlain [9] extends the approach to a more flexible specification of α as a function of x.

[14] The core (step 1 through 5) of the panel probit routine of Algorithm 5 requires only about 20 lines of code in Matlab.

Algorithm 5 Panel random effects probit

1. Draw $w_{it}|\beta, \alpha, \tau^2$ for all i and t:

   (a) for $y_{it} = 1$:

   $$w_{it}|\beta, \alpha \sim \mathcal{LTN}\left(x_{it}\beta + \alpha_i, 1\right)$$

   (b) for $y_{it} = 0$:

   $$w_{it}|\beta, \alpha \sim \mathcal{UTN}\left(x_{it}\beta + \alpha_i, 1\right)$$

2. Draw $\beta|w, \alpha, \tau^2$ from a Bayesian regression of $w - \alpha$ on x_{it}, with Normal priors on β and known variance $1 - \tau^2$:

   $$\beta|w \sim \mathcal{N}\left((X'X + A)^{-1}(X'(w - \alpha) + A\mu), \left(1 - \tau^2\right) \cdot (X'X + A)^{-1}\right)$$

   where $w - \alpha$is the stacked vector $\{w_{it} - \alpha_i\}$ across i and t, corresponding to the matrix X.

3. Draw $\alpha|\beta, w, \tau^2$ from a Bayesian regression of $w - X\beta$ on a $NTxN$ matrix of firm dummies D , using a $\mathcal{N}\left(0, I_N \cdot \left(1 - \tau^2\right)/\tau^2\right)$ prior:

   $$\alpha|\beta, w, \tau^2 \sim \mathcal{N}\Big(\left(D'D + I_N \cdot \left(1 - \tau^2\right)/\tau^2\right)^{-1} \cdot D'(w - X\beta), \left(1 - \tau^2\right)$$

   $$\cdot \left(D'D + I_N \cdot \left(1 - \tau^2\right)/\tau^2\right)^{-1}\Big)$$

4. Draw $\tau^2|\alpha, \beta, w$, using an $\mathcal{IG}\left(a, b\right)$ prior:

   $$\tau^2|\alpha, \beta, w \sim \mathcal{IG}\left(a + N, b + \sum_{i=1}^{N} \alpha_i^2\right)$$

5. Go back to step 1, repeat.

estimate by Maximum Likelihood, consider the likelihood function:

$$L = \prod_{i=1}^{N} \int_{-\infty}^{\infty} \left[\prod_{t=1}^{T} \Phi \left((2y_{it} - 1) \cdot \frac{x_{it}\beta + \alpha_i}{\sqrt{1-\tau^2}} \right) \right] \phi \left(\frac{\alpha_i}{\tau} \right) d\alpha_i$$

where, as before, $\phi(\cdot)$ is the pdf of the standard Normal distribution, and $\Phi(\cdot)$ is the cdf. The term in square brackets is the standard probit likelihood, conditional on α_i. Due to the non-linearity of $\Phi(\cdot)$, the expectation over α cannot be solved analytically, so numerical methods are required to evaluate the integral. In order to calculate the likelihood for one set of parameters, we need to evaluate N unidimensional integrals (in addition to the integrals required to evaluate the standard Normal cdf in the inner term of the likelihood). This can be done quite efficiently by Gauss-Hermite quadrature (this is how Stata estimates this model). However, even with small changes to the model the integral becomes of high dimension, at which point quadrature quickly loses its effectiveness (even for as few as five dimensions [98]). For example, allowing for auto-correlation in the η_{it}'s, the inner term, $\prod_{t=1}^{T} \Phi \left((2y_{it} - 1) \cdot \frac{x_{it}\beta + \alpha_i}{\sqrt{1-\tau^2}} \right)$, becomes a T-dimensional integral with no analytical expression.[15] Allowing instead for cross-correlation (but no auto-correlation) in the error terms, rewrite the likelihood as

$$\prod_{t=1}^{T} \int f\left(y_{1t} \dots y_{Nt}\right) g\left(\alpha_1 \dots \alpha_N\right) d(\alpha_1 \dots \alpha_N)$$

where $f(\cdot)$ is the joint likelihood of $y_{1t} \dots y_{Nt}$ conditional on the α's, and $g(\cdot)$ is the joint probability density of the RE's. Even conditional on the α's, the N-dimensional joint likelihood $f(\cdot)$ has no analytical solution, and on top of that one needs to integrate over the distribution of the RE's. The classical alternative to quadrature, Simulated Maximum Likelihood, thus requires a large simulation exercise to evaluate the likelihood, covering the entire joint distribution of all latent variables. This simulation has to be repeated for every guess of the parameters vector. MCMC on the other hand switches between drawing new parameters and drawing the latent states. The integration over the distribution of the latent states only needs to be done once, after the simulation is finished. This speeds up estimation considerably. For example, Jeliazkov and Lee [54] extend MCMC algorithm 5 and estimate a random effects panel probit model in which the η_{it} are serially correlated, and apply it to women's labor force participation. The estimation problem extends to many related models for which the likelihood is non-linear in the parameters, and Algorithm 5 can be adapted to estimate these models as well. For

[15] Note that the problem of integration with auto-correlated errors is closely related to the estimation of the dynamic selection model of section 3.3.

example, Bruno [7] develops an algorithm for a panel RE Tobit model. Such models have wide applicability in corporate finance for the same reasons as mentioned above: many standard variables, such as leverage, investment, or the decision to fire a CEO, are of a binary or truncated nature, and fixed or random effects help control for time-invariant unobserved heterogeneity in a panel data setting. Another useful extension is to deal with unbalanced panel data by combining Algorithm 5 with the randomly missing data Algorithm 2.

4.2 Panel data with selection/attrition

In this section I combine the Heckman model from section 3.2 with the panel RE model from the previous section. In other words, I allow for non-random missing data in a random effects panel model. This model is useful for controlling for non-random attrition, for example firms disappearing through bankruptcy or merger/acquistion. Despite the wide range of potential applications, no canned estimators are currently available in popular software packages.[16]

The model is

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}\beta + \delta \cdot \eta_{it} + \sigma_\xi \cdot \xi_{it} \\ w_{it} &= \theta_i + z_{it}\gamma + \eta_{it} \end{aligned}$$

where δ and σ_ξ are as defined in section 3.2. The random effects are iid, $\alpha_i \sim \mathcal{N}\left(\mu, \tau^2\right)$, and $\theta_i \sim \mathcal{N}\left(\kappa, \omega^2\right)$. The error terms are iid, $\eta_{it} \sim \mathcal{N}\left(0, 1-\omega^2\right)$, and $\xi_{it} \sim \mathcal{N}\left(0, \sigma_\xi^2\right)$, and are uncorrelated with each other and with the random effects. As in the standard Heckman model in section 3.2, selection enters through $\delta \neq 0$. Hausman and Wise [42] were the first to consider this model. Other Maximum Likelihood approaches have been developed by Ridder [87], Nijman and Verbeek [81] and Vella and Verbeek [100]. These models impose strong assumptions and are computationally burdensome as they require the evaluation of multiple integrals. To circumvent the computational burden, two-step estimators are typically employed, but these understate the standard errors of the parameters in the observation equation by not accounting for the estimation error in the inverse Mills ratios. MCMC is useful in these models because it is a computationally more efficient estimation method, and it allows for correct inference taking into account all sources of estimation error.

Algorithm 6 shows the MCMC procedure, which is essentially the standard Heckman

[16] Running a panel Heckman in Stata and using the "cluster" option to allow for random effects does not lead to the same result, as the error clustering is not true Maximum Likelihood and does not allow for random effects in the selection equation.

Algorithm 6 Panel random effects with selection

1. Draw $w_{it}, y_{it}^*|\alpha, \theta, \beta, \gamma, \delta, \sigma_\xi, \mu, \tau^2, \kappa, \omega^2$

   (a) for y_{it} observed:

   $$w_{it} \sim \mathcal{LTN}\left(\theta_i + z_{it}\gamma + \rho\sqrt{1-\omega^2} \cdot \left[\frac{y_{it} - \alpha_i - x_{it}\beta}{\sqrt{\delta^2 + \sigma_\xi^2}}\right], \left(1-\rho^2\right) \cdot \left(1-\omega^2\right)\right)$$

   where $\rho = \delta/\sqrt{\delta^2 + \sigma_\xi^2}$.

   (b) for y_{it} not observed:

   $$\begin{aligned} w_{it} &\sim \mathcal{UTN}\left(\theta_i + z_{it}\gamma, 1-\omega^2\right) \\ y_{it}^*|w_{it} &\sim \mathcal{N}\left(\alpha_i + x_{it}\beta + \delta\left[w_{it} - \theta_i - z_{it}\gamma\right], \sigma_\xi^2\right) \end{aligned}$$

2. Draw $\beta, \gamma|w, y^*, \alpha, \theta, \delta, \sigma_\xi, \mu, \tau^2, \kappa, \omega^2$ from a Bayesian Seemingly Unrelated Regression of $[y-\alpha; w-\theta]$ on $[X; Z]$, with Normal priors on β and γ and known covariance matrix Ω as in Algorithm 4 Step 2, where

   $$\Omega = \begin{bmatrix} \sigma_\xi^2 + \delta^2 & \delta\sqrt{1-\omega^2} \\ \delta\sqrt{1-\omega^2} & 1-\omega^2 \end{bmatrix} \otimes I_N$$

3. Draw $\alpha|w, y^*, \theta, \beta, \gamma, \delta, \sigma_\xi, \mu, \tau^2, \kappa, \omega^2$ from a Bayesian regression of $y - X\beta$ on a $NTxN$ matrix of firm dummies D , using a $\mathcal{N}\left(\mu, I_N \cdot \left(\delta^2 + \sigma_\xi^2\right)/\tau^2\right)$ prior.

4. Draw $\theta|w, y^*, \alpha, \beta, \gamma, \delta, \sigma_\xi, \mu, \tau^2, \kappa, \omega^2$ from a Bayesian regression of $w - Z\gamma$ on a $NTxN$ matrix of firm dummies D , using a $\mathcal{N}\left(\kappa, I_N \cdot \left(1-\omega^2\right)/\omega^2\right)$ prior.

5. Draw $\mu, \tau^2|w, y^*, \alpha, \theta, \beta, \gamma, \delta, \sigma_\xi, \kappa, \omega^2$ from a Bayesian regression of α on a constant, using Normal-IG priors (see Algorithm 1).

6. Draw $\kappa, \omega^2|w, y^*, \alpha, \theta, \beta, \gamma, \delta, \sigma_\xi, \mu, \tau^2$ from a Bayesian regression of θ on a constant, using Normal-IG priors (See Algorithm 1)

7. Draw $\delta, \sigma_\xi^2|w, y^*, \alpha, \theta, \beta, \gamma, \mu, \tau^2, \kappa, \omega^2$ from a Bayesian regression of $y - \alpha - X\beta$ on $w - \theta - Z\gamma$, with Normal-IG priors (see Algorithm 1).

8. Go back to step 1, repeat.

model of Algorithm 4, augmented with the RE components as in Algorithm 5. Dropping Steps 3 through 6 and setting μ, κ, τ and ω to zero collapses the algorithm back to the standard Heckman model. Similarly, forcing $\rho = 0$ collapses the algorithm down to a standard RE panel regression without selection. Although Algorithm 6 is slightly longer than the algorithms shown thus far, it is not much more complicated as it is still essentially a sequence of regressions.

To illustrate the importance of selection effects in corporate finance, I run a RE panel regression of quasi-market leverage (defined as the book value of debt divided by the book value of debt plus the market value of equity) on lagged profitability (operating income divided by the book value of assets), tangibility (net property plant and equipment divided by book assets), market-to-book ratio (market value divided by book value of equity) and the natural logarithm of book assets. Regressions of this nature are quite common in the literature (e.g. [66]), although researchers typically use fixed effects rather than random effects. I use data for a random sample of 1,000 firms from Compustat between 1950 and 2010, for a total of 11,431 firm-years. The first column in Table 3 shows the standard GLS estimates of the regression coefficients. To gauge the effect of the priors and the additional distributional assumptions of the MCMC algorithm in this example, the second column of the table reports the MCMC estimates of the same regression model, ignoring the selection issue (i.e. forcing $\rho = 0$). The coefficient estimates are very close, suggesting that MCMC indeed replicates the standard GLS regression when ignoring selection. The random effects are important as they account for 57% of the variance (result not reported in the table).

An important concern is that leverage is missing for 1,572, or 14%, of firm-years. If observations are not missing at random, for example if firms drop out of the sample due to a merger or a bankruptcy, this can result in misleading estimates. Unreported results reveal that the missing firm-years are characterized by higher profitability, higher tangibility and lower market-to-book ratios, suggesting that they are indeed different in observable respects. This is in principle not a problem, unless there is also selection on *unobservable* variables, which would manifest itself as a non-zero correlation between the error terms in the selection and observation equations, ρ. The last column of Table 3 shows the estimates from MCMC Algorithm 6. The posterior mean of the correlation ρ is 0.367, the first percentile of the posterior distribution is 0.237 and the 1/10th percentile is 0.070. This finding indicates that ρ is not zero and there is indeed non-random selectivity on unobserved variables. Moreover, the selection correction has a large impact on parameter estimates. The coefficient on profitability drops from -0.186 in the GLS to -0.237 after correcting for selection, a change of about 25%. Moreover, the coefficient on tangibility

Dependent variable: Quasi-market leverage			
	GLS	MCMC	
		No selection correction	Selection correction
Operating income / assets	-0.183	-0.186	-0.237
	(0.015)	(0.015)	(0.017)
Tangibility	0.209	0.207	0.101
	(0.016)	(0.016)	(0.021)
Market-to-book	-0.029	-0.029	-0.028
	(0.002)	(0.002)	(0.002)
Log(assets)	0.020	0.020	0.019
	(0.002)	(0.002)	(0.002)
ρ	-	-	0.367
	-	-	(0.049)

Table 3: **Random effects panel regressions.**
Random effects panel regression estimates of quasi-market leverage (defined as the book value of debt divided by the book value of debt plus the market value of equity) on profitability (operating income divided by the book value of assets), tangibility (net property plant and equipment divided by book assets), market-to-book ratio (market value divided by book value of equity) and the natural logarithm of book assets, based on a random sample of 1,000 firms spanning 11,431 firm-years from Compustat between 1950 and 2010. All explanatory variables are lagged by one period. The first column uses the standard GLS method to estimating a random effects panel model. The estimates in the "No Selection" column show the same random effects model, but using MCMC (Algorithm 6 with the correlation between the error terms, ρ, forced to zero). The column labeled "Selection" uses the panel random effects with attrition model in MCMC Algorithm 6 to correct for sample selection. The MCMC estimates use 1,000 burn-in cycles followed by 10,000 cycles to sample the posterior distribution. Standard errors for the GLS estimates and posterior standard deviations for the MCMC estimates are in brackets.

roughly drops in half, from 0.207 to 0.101. The estimates of market-to-book and firm size are only marginally affected by the selection issue. Note also the well-known result that the estimate of residual variance (and hence standard errors) is biased downwards when ignoring selection [43], hence the larger posterior standard deviations of the coefficients in the selection correction model. For brevity I do not report the coefficients of the selection equation, although these could be interesting in their own right as they convey information about the reasons for selection.

It is important to note that the results in Table 3 should be taken as suggestive only. There are other observable variables that can be included in the observation and selection equations that may alleviate the omitted variable problem. Moreover, I include the same variables in the selection equation as in the observation equation, so the model is identified from distributional assumptions only. A more thorough analysis requires an instrument that changes the probability of observing leverage but does not drive leverage itself. In other words, one needs a variable in the selection equation that does not appear in the observation equation. This requires delving deeper into the economic reasons for selection. For example, the fact that more profitable firms are more likely to disappear

from the data suggests that the main source of attrition is mergers and acquisitions rather than bankruptcy. One would then look for an instrument that drives M&A but not leverage. Other possible extensions are to use two selection equations to separately model bankruptcy and M&A as different reasons for attrition, essentially producing a joint model of capital structure, M&A and bankruptcy, or to use the Korteweg and Sørensen [61] approach and model leverage as an AR(1) process. Note also that the above example, given the use of random effects, only considers selection in the time series. There could be important cross-sectional sample selection issues over and above the time-series selection problem.

It is fairly straightforward to generalize Algorithm 6 to make the dependent variable a binary (probit) or truncated (Tobit) variable. For example, Hamilton [39] specifies a RE panel Tobit selection model with t-distributed error terms and uses MCMC to estimate the effect of HMO choice on healthcare costs. A different extension to the model is in Cowles, Carlin, and Connett [20] who estimate a panel selection model allowing for two observation equations. Selection is multinomial and the observation equations are observed at different cutoffs of the selection variable. They apply their model to a longitudinal clinical trial measuring the effect of smoking cessation paired with an inhaler treatment on lung function. The selection problem is that patients endogenously do not show up for follow-up visits, or show up but fail to return canisters with inhaler.

5 Structural models

Structural models have many potential applications in empirical corporate finance, and have been used in areas such as capital structure (e.g. [47, 48, 95]) and corporate governance (e.g. [97]). These models are typically estimated by simulated method of moments (SMM) or, in some cases, Simulated Maximum Likelihood (SML, see e.g. [6, 82]). In this section I will illustrate the benefits of estimating these models by MCMC, especially models with latent state variables.

Consider Merton's [77] model of the firm in state-space representation:

$$\begin{aligned} v_t &= v_{t-1} + \mu + \sigma\varepsilon_t \\ E_t &= P\left(v_t, \sigma; \theta\right) + \eta_t \end{aligned}$$

The state variable of the model is the natural logarithm of the market value of the firm's assets, v_t, which follows a random walk with drift μ. The shocks to firm value, ε_t, have a standard Normal distribution. The observed market value of equity, E_t, is a function of

v_t, volatility, σ, and other parameters, θ. In Merton's model the pricing function, $P(\cdot)$, is the Black-Scholes European call option function, and the parameters vector θ consists of the maturity of debt (time-to-expiration), the face value of debt (strike price), and the risk-free interest rate, all of which I will assume are observed. The pricing error, η_t, allows for unmodeled features such as market microstructure noise [51, 60] that may force observed prices away from the model.[17]

I assume for now that, in addition to θ, the time series of firm value, $v^T = \{v_1 \dots v_T\}$ as well as the equity values are all observed. The inference problem is then to obtain the posterior distribution $p\left(\sigma^2, \mu | E^T, v^T\right)$. As usual, an MCMC algorithm cycles between drawing from the complete conditionals, $p\left(\mu|\sigma^2, E^T, v^T\right)$ and $p\left(\sigma^2|\mu, E^T, v^T\right)$. The first distribution is a simple Bayesian regression and poses no problems. The second distribution is more tricky because σ is present in both equations of the state-space, and although the conditional posterior can be evaluated without much trouble, it is not a known distribution that is easily sampled. The solution to this problem is an accept-reject type algorithm known as Metropolis-Hastings.[18] Denote by ς the current draw of $\left\{\sigma^2\right\}^{(g)}$ in cycle g of the MCMC algorithm. To sample the next draw, $\left\{\sigma^2\right\}^{(g+1)}$, one proceeds as follows:

1. Draw s from a proposal density $f(s;\varsigma)$.
2. Compute $\alpha = \min\left(1, \frac{p\left(s|\mu,E^T,v^T\right)\cdot f(\varsigma;s)}{p\left(\varsigma|\mu,E^T,v^T\right)\cdot f(s;\varsigma)}\right)$.
3. Set $\left\{\sigma^2\right\}^{(g+1)} = \begin{cases} s & \text{w/ probability } \alpha \\ \varsigma & \text{w/ probability } 1-\alpha \end{cases}$.

Note that we use the latest draw for μ when computing α. The proposal density is key for the performance of the algorithm. Ideally, it is a density that is close to the target density, $p\left(\sigma^2|\mu, E^T, v^T\right)$, yet easy to sample from. The proposals can be iid (commonly known as "independence Metropolis") or proposals may depend on the current draw of σ^2. A popular example of the latter type is the "random walk Metropolis" in which s equals the current draw, ς, plus a random increment. Note that if we are able to sample from the target density, i.e. $f(\cdot) = p(\cdot)$, then we are back to a Gibbs sampler step: $\alpha = 1$ and we always accept the proposal. In this sense we can think of the Metropolis-Hastings algorithm as a generalization of the Gibbs sampler. As in the Merton model example, an

[17] An alternative motivation for the pricing errors is simply to avoid a stochastic singularity problem in the sense that the model makes predictions about more observable variables than there are structural shocks and hence could be rejected with the observation of as few as two time periods.

[18] The Metropolis-Hastings algorithm is derived from the detailed balance condition of the stationary distribution of the Markov chain, which ensures that the chain will be reversible. I will not go into the theoretical details here, but instead refer the reader to Johannes and Polson [55], Robert and Casella [88], or Rossi et al. [89] for a comprehensive treatment.

MCMC algorithm can be a mixture of Gibbs steps and Metropolis-Hastings steps.

It is important to "tune" the proposal density to get reasonable performance from the sampler. Clearly an acceptance rate near zero is bad because the sampler will be slow to converge to, and explore, the posterior distribution due to the many rejected proposals. An acceptance rate near 100% may seem great at first sight, but one may worry that the chain is moving very slowly because the incremental steps in a random walk Metropolis are too small, or because the tails of the proposal are too thin relative to the target distribution in an independence Metropolis sampler. Either way, we may be undersampling an important part of the posterior distribution, leading to poor inference. There is no generic theoretical advice that can be given as to how to best pick a proposal distribution. Typical advice in practice is to aim for a 50-80% acceptance rate, but is should be noted that the acceptance rate alone does not determine whether the sampler does a good job. For more details on the Metropolis-Hastings algorithm and tuning, see Johannes and Polson [55], Robert and Casella [88], or Rossi et al. [89].

In many structural models inference is complicated further by the fact that the underlying state variable is not observed (e.g. [47, 48, 65]). MCMC deals with this by augmenting the posterior distribution with the state variable, $p\left(\sigma, \mu, v^T | E^T\right)$, which can subsequently be integrated out to obtain the marginal distribution of the model parameters. Conveniently, the complete conditionals for μ and σ do not need any modification (they are already conditioned on a draw of v^T), so the only extra step that is required is to sample from the conditional distribution of the state variable $p\left(v^T | \mu, \sigma, E^T\right)$. In the Merton model a linear Kalman filter provides us with the distribution of $v_1|E_1$ through $v_T|E_T$, dropping the conditioning on the parameters for ease of exposition. Since we need to draw from $v_1|E^T$ through $v_T|E^T$, one more step is required: smoothing. This essentially involves running the Kalman filter again, but going backwards starting at time T, and sampling the v_t's as we go along. This procedure is called "Forward Filtering, Backwards Sampling" (FFBS) [8, 31].

Korteweg and Polson [60] describe the MCMC algorithm for a structural model of the firm in detail, with FFBS and Metropolis-Hasting sampling for σ^2. They estimate Leland's [64] model for a panel of firms and compute corporate bond credit spreads, taking into account the effect of parameter and latent state uncertainty. Using the robust pricing framework of Hansen and Sargent [41], the bond price under uncertainty, P_t, is the expectation of the model price over the distribution of latent state and parameters

given the data up to time t:

$$\begin{aligned} P_t &= \mathbb{E}_{\sigma,v_t|E^t}\left(B(v_t,\sigma;\theta)\right) \\ &\approx \frac{1}{G}\sum_{i=1}^{G} B\left(v_t^{(i)},\sigma^{(i)};\theta\right) \end{aligned}$$

where $B(\cdot)$ is the model's bond pricing function. Bayesian methods are very well suited for problems of learning and uncertainty, since the posterior distribution captures the degree of parameter and state uncertainty based on the observed data. The second line shows the approximation to the bond price based on G cycles of the MCMC algorithm (after dropping the initial burn-in cycles). The expectation in the top line becomes a simple average over the draws of the algorithm. The concavity of the bond pricing formula in v and σ results in larger credit spreads for bonds compared to estimates from standard methods (e.g. [28]).

Apart from the usefulness for learning and uncertainty problems, MCMC methods have the same advantage over SMM and SML in speed of computation as described in section 4.1: instead of simulating the latent state variables for each set of parameters as is required in SMM/SML, the MCMC algorithm bounces between parameter draws and draws of the state vector, leading to faster convergence. Moreover, the MCMC and other likelihood-based estimates incorporate all available information, unlike SMM which only considers a set of moments chosen by the researcher and therefore results in less powerful tests. However, this efficiency gain does come at the expense of making stronger distributional assumptions on the observation error, η (see Korteweg and Lemmon [59] for a detailed discussion of structural model testing). Finally, as shown earlier, MCMC gives correct small sample inference and is amenable to computing non-linear functions of parameters and states. This is a particularly useful feature for structural models where the observables are often non-linear in both the parameters and states.

The MCMC algorithm for structural models can be extended in various important directions, including but not limited to panel data, autocorrelation in the error structure, handling missing data, and adding stochastic volatility and jumps to the state process, as well as more flexible observation error distributions, for example through mixtures of Normals. The inherent modularity of MCMC makes this task more convenient to handle than with classical methods.

There are other applications of state space models outside of strict structural modeling. One example is Korteweg [58], who uses MCMC to estimate the net benefits to leverage from a state space model in which the state vector is composed of individual firms' unlevered asset values and industry asset betas. The observations are the market

values of firms' debt and equity and their riskiness (i.e. betas) with respect to the market portfolio. Under the identifying assumption that all firms within the industry share the same (unlevered) asset beta, Korteweg identifies the present value of the net benefits to leverage as a function of leverage and other firm characteristics such as profitability, market-to-book ratio, and asset tangibility. The MCMC algorithm for this model is similar to the algorithms described above, employing the Kalman filter and FFBS to integrate out the latent states.

6 Extensions and other applications

In this section I will highlight some further potential applications of MCMC methods in corporate finance, focusing on cases which are difficult to estimate using classical methods.

Hierarchical models are a particularly useful type of model that has to date seen little application in corporate finance. Hierarchical models were introduced in section 4.1 in the context of estimating random effects in a panel probit model, but the basic concept of the hierarchical model has much broader applicability. For example, consider a model of the decision to issue equity. The typical approach would be to run a probit or hazard model that specifies the probability of issuing equity as a function of a number of covariates, most importantly the financing deficit, one of the key variables in the pecking order theory of capital structure. However, a common concern is that the financing deficit and the decision to issue equity are jointly determined in the sense that they are both driven by (unobserved) investment opportunities. Lacking a good instrument or natural experiment, one example of a hierarchical modeling solution to this problem is to model the deficit conditional on an underlying (latent) investment opportunity variable, and specify the probability of issuing equity conditional on both the deficit and the latent variable, potentially allowing for correlation between the error terms.[19]

It is also useful to add hierarchical model features to structural models. For example, it would be interesting to analyze the cross-sectional distribution of firms' bankruptcy costs, or the distribution of CEO talent in structural models of the firm. However, there is usually not enough data to reliably pin down the estimate on a firm by firm basis. One way to overcome this issue is to use a hierarchical setup and specify a firm's bankruptcy cost or a CEO's talent as being generated from a particular distribution, using the observed data to estimate the mean and variance of this distribution, as was done in the random effects model.

[19] Of course an instrument that shocks the deficit (but not the equity issuance probability) will help the identification of the model. The intuition is similar to the argument for having an instrument in the Heckman selection model.

Bayesian methods are also very useful for duration (hazard) models. For example, one could model the probability of attrition in section 4.2 through a hazard rate instead of a probit model. This would do a better job of capturing the essence of liquidiation or an acquisition in the sense that, unlike in the probit model, a firm could not come back to life the next period. Li [68] estimates the duration of Chapter 11 bankruptcy spells from a Bayesian perspective, using an approximation to the posterior density, accounting for parameter and model uncertainty as well as providing correct sample properties based on the small dataset. Horny, Mendes and Van den Berg [49] estimate a hazard model of job duration with worker and firm-specific random effects. They use random effects because fixed effects are not feasible due to right-censoring of the unemployment spells, and because the random effects allow them to decompose the variation of job durations into the relative contributions of workers' and firms' characteristics. In addition, they allow for multiple job spells per worker and for correlation between the worker and firm random effects. They estimate the model by MCMC because the joint dependence of the random effects makes classical likelihood-based estimation very difficult for both computational speed and the difficulty of finding the appropriate asymptotic distribution in order to calculate standard errors. Chen, Guo and Lin [12] develop a model of switching hazard models which they use to estimate the probability of IPO withdrawal in relation to its subsequent survival hazard, which in turn depends on the IPO decision. They use MCMC because of the computational advantage in both estimation and performing model selection and cross-validation, as well as the small-sample inference properties. Fahrmeir and Knorr Held [29] develop a non-parametric MCMC method for a duration model with multiple outcomes, such as unemployment ending in a full-time vs a part-time job. Such a model could, for example, be applied to firms' survival spells ending in either bankruptcy or an acquisition, or CEO's tenure spells ending in a forced or voluntary retirement.

Another area where MCMC can be applied is count data. These models are related to duration models, but instead of a time spell the dependent variable is a non-negative integer that counts the number of events that have occurred within a given time period. Examples include the number of takeover bids received by a target firm [53], the number of failed banks [23] or the number of defaults in a portfolio of assets [57]. Chib, Greenberg and Winkelmann [17] develop an MCMC algorithm for panel count data models with random effects. They argue that Maximum Likelihood is not a viable estimator for this model due to the presence of random effects paired with the non-linearity of count data. Chib and Winkelmann [18] generalize this model to allow for multivariate correlated count data, represented by correlated latent effects. Munkin and Trivedi [79] embed a count model in a self-selection model with two correlated outcome equations, one of which is a count and

the other a continuous variable. The authors strongly motivate their MCMC approach by the computational diffculties encountered when attempting to estimate the model by SML. Deb, Munkin and Trivedi [24] extend this selection model by allowing the entire outcome function to be different among the treated and untreated groups, essentially turning the model into a switching regression with count outcomes. They apply their MCMC estimation to separate the incentive and selection effects in private insurance on the number of doctor visits, using a multiyear sample of the U.S. adult non-Medicare population. This could be applied in corporate finance, for example, for identifying the incentive and selection effects in public versus private firms in undertaking acquisitions, using the count of M&A as the dependent variable.

A class of models that is very closely related to the switching regressions of section 3.4, and is also related to the state space models of section 5 is the set of hidden Markov models (HMM). Unlike switching regressions, there are typically no covariates that drive the selection into a specific state, but they generally allow for multiple states (in some cases a continuum of states) of the system that switch according to a Markov process. In economics, HMM are usually applied to regime switching settings, where the economy may switch between expansion and recession. Albert and Chib [2] argue that the two-step Maximum Likelihood approach to estimating HMM of Hamilton [40] does not give correct standard errors as the uncertainty about the parameters in the first step is not incorporated in the second step. In addition, the ML approach does not provide a complete description of the likelihood, such as bimodality or asymmetry. Albert and Chib develop an MCMC algorithm that deals with both issues. McCullogh and Tsay [76] use MCMC to estimate an HMM that is very close to the switching regressions of section 3.4 and apply it to GNP data. Ghysels, McCullogh and Tsay [36] estimate a non-linear regime switching model and apply their MCMC estimator to two examples, one using housing starts data while the other employs industrial production data. They argue that the MCMC approach is particularly suitable to their model because classical estimation of periodic Markov chain models often results in parameter estimates at the boundary.

Recent work has started exploring the uses of MCMC for Instrumental Variables, with some success (see Sims [92] for a discussion and further references). In particular when the error distributions are non-Normal, the Bayesian estimates may be more efficient than classical methods [19]. On a different note, two recent papers by Davies and Taillard [21, 22] propose an MCMC approach to dealing with omitted variables that does not use an instrument at all, but instead achieves identification by assuming that any common variation in the residuals is due solely to the omitted variables.

Finally, thanks to the modularity of MCMC, models can be mixed and extended in

various ways that are quite straightforward. For example, one can deal with missing data by adding the Algorithm 2 steps to any of the other algorithms. Error clustering can be handled by adding random effects to the models. Last but not least, one can deal with heteroskedasticity and non-Normal error terms by using mixtures of Normals (see Diebolt and Robert [25] and Geweke [35] for MCMC mixture models, and Chen and Liu [13] for Kalman filters with mixtures of Normals, and Korteweg and Sørensen [61] for an application). These mixture models can be added to MCMC algorithms with relative ease, are very flexible in generating both skewness and kurtosis in the error distributions, and make the MCMC estimates more robust to outliers.

7 Conclusion

With the current trend towards more complex empirical models in the corporate finance literature, Markov Chain Monte Carlo methods provide a viable and attractive means of estimating and evaluating models where classical methods such as least-squares regression, GMM and Maximum Likelihood and their simulated counterparts are difficult or too computationally demanding to apply. In particular, this includes non-linear models with many latent variables that require high-dimensional integration to evaluate the likelihood, or models that have a hierarchical structure. Examples of such models include panel limited dependent variable models, matching and other self-selection models, and structural models of the firm. The potential application of these types of models in corporate finance is vast, including such diverse areas as capital structure, financial intermediation, bankruptcy, and corporate governance.

The core feature of the method is the Hammersley-Clifford theorem, which breaks up the problem into its complete conditional distributions. These are usually relatively easy to sample from, requiring no more than standard regression tools. Another benefit of the MCMC approach is that is is modular, so that, for example, one can add a missing data module to a panel probit algorithm with relative ease. Moreover, the method allows for exact small-sample inference of parameters and non-linear functions of parameters (the latter being helpful, for example, when calculating marginal effects in a probit or logit model), and does not require optimization algorithms, simulated annealing or other methods that can make Maximum Likelihood and GMM cumbersome to use.

Every introductory text necessarily has to focus on certain ideas at the expense of others, and I have chosen to focus on applications rather than to discuss some of the more technical details of the MCMC method, the role of priors, and convergence diagnostics. Most introductory textbooks thoroughly discuss these and other topics, and I refer the

interested reader to Rossi et al. [89] and Johannes and Polson [55] for a particularly lucid treatise and further reading.

This chapter has only touched the tip of the iceberg of possibilities that MCMC has to offer for empirical corporate finance research, and I hope to have convinced the reader that the potential benefits of the method are plentiful. Given that learning MCMC does come with some fixed cost, I hope that this chapter and the accompanying code samples help to lower the cost of adoption, and inspire and motivate corporate finance researchers to dive deeper into MCMC methods. With time, hopefully this methodology will become part of the standard toolkit in finance, as it already is in many other areas of scientific inquiry.

References

[1] Acharya, S. (1988). A generalized econometric model and tests of a signalling hypothesis with two discrete signals. *Journal of Finance*, **43**, 412–429.

[2] Albert, J. and Chib, S. (1993). Bayes inference via Gibbs sampling of autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, **11**, 1–15.

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

[4] Berger, J.O. and Delampady, M. (1987). Testing precise hypotheses. *Statistical Science*, **2**, 317–335.

[5] Billet, M.T. and Xue, H. (2007). The takeover deterrent effect of open market share repurchases. *Journal of Finance*, **62**, 1827–1850.

[6] Bruche, M. (2007). Estimating structural models of corporate bond prices. Working paper, CEMFI Madrid.

[7] Bruno, G. (2004). Limited dependent panel data models: A comparative analysis of classical and Bayesian inference among econometric packages. Working paper, Bank of Italy.

[8] Carter, C.K. and Kohn, R.J. (1994). On Gibbs sampling for state space models. *Biometrika*, **81**, 541–553.

[9] Chamberlain, G. (1984). Panel data. In *Handbook of Econometrics* (ed. Z. Griliches and M. Intriligator). Elsevier, Amsterdam: North Holland.

[10] Chen, J. (2010). Two-sided matching and spread determinants in the loan market. Working paper, UC Irvine.

[11] Chen, L. and Zhao, X. (2007). Mechanical mean reversion of leverage ratios. *Economics Letters*, **95**, 223–229.

[12] Chen, R., Guo, R.-J., and Lin, M. (2010). Self-selectivity in firm's decision to withdraw IPO: Bayesian inference for hazard models of bankruptcy with feedback. Working paper, Rutgers University.

[13] Chen, R. and Liu, J.S. (2000). Mixture Kalman filters. *Journal of the Royal Statistical Society Series B*, **62**, 493–508.

[14] Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, **51**, 79–99.

[15] Chib, S. and Greenberg, E. (1998). Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.

[16] Chib, S., Greenberg, E., and Chen, Y. (1998). MCMC methods for fitting and comparing multinomial response models. Working paper, Washington University in St. Louis.

[17] Chib, S., Greenberg, E., and Winkelmann, R. (1998). Posterior simulation and Bayes factors in panel count data models. *Journal of Econometrics*, **86**, 33–54.

[18] Chib, S. and Winkelmann, R. (2001). Markov Chain Monte Carlo analysis of correlated count data. *Journal of Business and Economic Statistics*, **19**, 428–435.

[19] Conley, T.G., Hansen, C.B., McCulloch, R.E., and Rossi, P.E. (2008). A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics*, **144**, 276–305.

[20] Cowles, M.K., Carlin, B.P., and Connett, J.E. (1996). Bayesian Tobit modeling of longitudinal ordinal clinical trial compliance data with nonignorable missingness. *Journal of the American Statistical Association*, **91**, 86–98.

[21] Davies, P. and Taillard, J. (2010). Omitted variables, endogeneity, and the link between managerial ownership and firm performanc. Working paper, University of Iowa and Boston College.

[22] Davies, P. and Taillard, J. (2011). Estimating the return to education: An instrument-free approach. Working paper, University of Iowa and Boston College.

[23] Davutyan, N. (1989). Bank failures as Poisson variates. *Economics Letters*, **29**, 333–338.

[24] Deb, P., Munkin, M.K., and Trivedi, P.K. (2006). Private insurance, selection, and health care use: A Bayesian analysis of a Roy-type model. *Journal of Business and Economic Statistics*, **24**, 403–415.

[25] Diebolt, J. and Robert, C.P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society Series B*, **56**, 363–375.

[26] Eckbo, B.E., Maksimovic, V., and Williams, J. (1990). Consistent estimation of cross-sectional models in event studies. *Review of Financial Studies*, **3**, 343–365.

[27] Efron, B. and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, **70**, 311–319.

[28] Eom, Y.H., Helwege, J., and Huang, J. (2004). Structural models of corporate bond pricing: An empirical analysis. *Review of Financial Studies*, **17**, 499–544.

[29] Fahrmeir, L. and Knorr Held, L. (1997). Dynamic discrete-time duration models. Working paper, University of Munich.

[30] Frank, M.Z. and Goyal, V.K. (2009). Capital structure decisions: Which factors are reliably important? *Financial Management*, **38**, 1–37.

[31] Fruhwirth-Schnatter, S. (1994). Data augmentation and dynamic linear models. *Journal of Time Series Analysis*, **15**, 183–202.

[32] Gelfand, A. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

[33] Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–511.

[34] Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

[35] Geweke, J. (2006). Interpretation and inference in mixture models: Simple MCMC works. Working paper, University of Iowa.

[36] Ghysels, E., McCulloch, R.E., and Tsay, R.S. (1998). Bayesian inference for periodic regime-switching models. *Journal of Applied Econometrics*, **13**, 129–143.

[37] Graham, B.S. and Hirano, K. (2011). Robustness to parametric assumptions in missing data models. *American Economic Review Papers & Proceedings*, **101**, 538–543.

[38] Griliches, Z. (1986). Economic data issues. In *Handbook of Econometrics, Volume 3* (ed. Z. Griliches and M. Intriligator), pp. 1466–1514. Elsevier, Amsterdam: North Holland.

[39] Hamilton, B. (1999). HMO selection and medicare costs: Bayesian MCMC estimation of a robust panel data Tobit model with survival. *Health Economics*, **8**, 403–414.

[40] Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**, 357–384.

[41] Hansen, L. and Sargent, T. (2010). Fragile beliefs and the price of uncertainty. *Quantitative Economics*, **1**, 129–162.

[42] Hausman, J.A. and Wise, D.A. (1979). Attrition bias in experimental and panel data: The Gary income maintenance experiment. *Econometrica*, **47**, 455–473.

[43] Heckman, J.J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Annals of Economic and Social Measurement*, **5**, 475–492.

[44] Heckman, J.J. (1979). Sample selection bias as a specification error. *Econometrica*, **47**, 153–161.

[45] Heckman, J.J. (1981). The incidental parameters problem and the problem of initial conditions in estimating a discrete time - discrete data stochastic process. In *Structural Analysis of Discrete Data with Econometric Applications* (ed. C. Manski and D. McFadden). MIT Press, Cambridge.

[46] Heckman, J.J. (1990). Varieties of selection bias. *American Economic Review*, **80**, 313–318.

[47] Hennessy, C.A. and Whited, T.M. (2005). Debt dynamics. *Journal of Finance*, **60**, 1129–1165.

[48] Hennessy, C.A. and Whited, T.M. (2007). How costly is external financing? Evidence from a structural estimation. *Journal of Finance*, **62**, 1705–1745.

[49] Horny, G., Mendes, R., and den Berg, G.J. Van (2009). Job durations with worker and firm specific effects: MCMC estimation with longitudinal employer-employee data. Working paper, IZA Insitute for the study of labor, Bonn.

[50] Hovakimian, A., Hovakimian, G., and Tehranian, H. (2004). Determinants of target capital structure: The case of dual debt and equity issues. *Journal of Financial Economics*, **71**, 517–540.

[51] Huang, S. and Yu, J. (2010). Bayesian analysis of structural credit risk models with microstructure noises. *Journal of Economic Dynamics and Control*, **34**, 2259–2272.

[52] Iliev, P. and Welch, I. (2010). Reconciling estimates of the speed of adjustment of leverage ratios. Working paper, UCLA.

[53] Jaggia, S. and Thosar, S. (1995). Contested tender offers: An estimate of the hazard function. *Journal of Business and Economic Statistics*, **13**, 113–119.

[54] Jeliazkov, I. and Lee, E.H. (2010). MCMC perspectives on simulated likelihood estimation. In *Advances in Econometrics* (ed. W. Green and R. Hill), Volume 26, pp. 3–39. Emerald Group Publishing Limited.

[55] Johannes, M. and Polson, N. (2012). *Computational Methods for Bayesian Inference: MCMC methods and Particle Filtering*. Unpublished manuscript.

[56] Kass, R.O. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

[57] Koopman, S.J., Lucas, A., and Schwab, B. (2010). Macro, industry and frailty effects in defaults: The 2008 credit crisis in perspective. Working paper, Tinbergen Institute.

[58] Korteweg, A. (2010). The net benefits to leverage. *Journal of Finance*, **65**, 2137–2170.

[59] Korteweg, A. and Lemmon, M. (2011). Structural models of capital structure: A framework for model evaluation and testing. Working paper, Stanford University.

[60] Korteweg, A. and Polson, N.G. (2010). Corporate credit spreads under uncertainty. Working paper, University of Chicago.

[61] Korteweg, A. and Sørensen, M. (2010). Risk and return characteristics of venture capital-backed entrepreneurial companies. *Review of Financial Studies*, **23**, 3738–3772.

[62] Korteweg, A. and Sørensen, M. (2011). Estimating loan-to-value and foreclosure behavior. Working paper, Stanford University.

[63] Lee, L.-F. (1979). Identification and estimation in binary choice models with limited (censored) dependent variables. *Econometrica*, **47**, 977–996.

[64] Leland, H. (1994). Bond prices, yield spreads, and optimal capital structure with default risk. Working paper, UC Berkeley.

[65] Leland, H. (1994). Risky debt, bond covenants and optimal capital structure. *Journal of Finance*, **49**, 1213–1252.

[66] Lemmon, M.L., Roberts, M.R., and Zender, J.F. (2008). Back to the beginning: Persistence and the cross-section of corporate capital structure. *Journal of Finance*, **63**, 1575–1608.

[67] Li, K. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics*, **85**, 387–400.

[68] Li, K. (1999). Bayesian analysis of duration models: an application to Chapter 11 bankruptcy. *Economics Letters*, **63**, 305–312.

[69] Li, K. and McNally, W. (2004). Open market versus tender offer share repurchases: A conditional event study. Working paper, University of British Columbia.

[70] Li, K. and Prabhala, N.R. (2007). Self-selection models in corporate finance. In *Handbook of Corporate Finance: Empirical Corporate Finance* (ed. B. Eckbo), Volume I, Chapter 2, pp. 37–86. Elsevier, Amsterdam: North Holland.

[71] Lindley, D.V. and Smith, A.F.M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

[72] Maddala, G.S. (1996). Applications of limited dependent variable models in finance. In *Handbook of Statistics* (ed. G. Maddala and G. Rao), Volume 14, pp. 553–566. Elsevier, Amsterdam: North Holland.

[73] Manski, C. (1989). Anatomy of the selection problem. *Journal of Human Resources*, **24**, 343–360.

[74] Manski, C. (1990). Nonparametric bounds on treatment effects. *American Economic Review*, **80**, 319–323.

[75] McCulloch, R.E., Polson, N.G., and Rossi, P.E. (2000). A Bayesian analysis of the multinomial probit model with fully identied parameters. *Journal of Econometrics*, **99**, 173–193.

[76] McCulloch, R.E. and Tsay, R.S. (1994). Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis*, **15**, 523–539.

[77] Merton, R.C. (1974). On the pricing of corporate debt: The risk structure of interest rates. *Journal of Finance*, **29**, 449–470.

[78] Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica*, **46**, 69–85.

[79] Munkin, M.K. and Trivedi, P.K. (2003). Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare. *Journal of Econometrics*, **114**, 197–220.

[80] Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, **16**, 1–32.

[81] Nijman, T. and Verbeek, M. (1992). Nonresponse in panel data: The impact on estimates of the life cycle consumption function. *Journal of Applied Econometrics*, **7**, 243–257.

[82] Nikolov, B., Morellec, E., and Schuerhoff, N. (2011). Corporate governance and capital structure dynamics. *Journal of Finance*, **forthcoming**.

[83] Obrizan, M. (2011). A Bayesian model of sample selection with a discrete outcome variable: Detecting depression in older adults. Working paper, Kyiv School of Economics.

[84] Park, M. (2008). An empirical two-sided matching model of acquisitions: Understanding merger incentives and outcomes in the mutual fund industry. Working paper, UC Berkeley.

[85] Prabhala, N.R. (1997). Conditional methods in event studies and an equilibrium justification for standard event-study procedures. *Review of Financial Studies*, **10**, 1–38.

[86] Pulvino, T. (1999). Effects of bankruptcy court protection on asset sales. *Journal of Financial Economics*, **52**, 151–186.

[87] Ridder, G. (1990). Attrition in multi-wave panel data. In *Panel data and labor market studies* (ed. G. R. J. Hartog and J. Theeuwes). Elsevier, Amsterdam: North Holland.

[88] Robert, C.P. and Casella, G. (2004). *Monte Carlo Statistical Methods* (2nd edn). Springer, New York.

[89] Rossi, P.E., Allenby, G.M., and McCulloch, R. (2005). *Bayesian Statistics and Marketing* (1st edn). John Wiley and Sons, Hoboken, NJ.

[90] Rubin, D. (1983). Some applications of Bayesian statistics to educational data. *The Statistician*, **32**, 55–68.

[91] Scruggs, J.T. (2007). Estimating the cross-sectional market response to an endogenous event: Naked vs. underwritten calls of convertible bonds. *Journal of Empirical Finance*, **14**, 220–247.

[92] Sims, C.A. (2007). Thinking about instrumental variables. Working paper, Princeton University.

[93] Sørensen, M. (2007). How smart is smart money? A two-sided matching model of venture capital. *Journal of Finance*, **62**, 2725–2762.

[94] Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the third Berkeley symposium.*

[95] Strebulaev, I.A. (2007). Do tests of capital structure theory mean what they say? *Journal of Finance*, **62**, 1747–1787.

[96] Tanner, M.A. and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–549.

[97] Taylor, L.A. (2010). Why are CEOs rarely fired? Evidence from structural estimation. *Journal of Finance*, **65**, 2051–2087.

[98] Train, K. (2003). *Discrete choice methods with simulation* (1st edn). Cambridge University Press, New York.

[99] van Hasselt, M. (2009). Bayesian inference in a sample selection model. Working paper, University of Western Ontario.

[100] Vella, F. and Verbeek, M. (1994). Two-step estimation of simultaneous equation panel data models with censored endogenous variables. Working paper, Tilburg University.